Best explanations, natural concepts, and optimal design

Igor Douven

Best Explanations, Natural Concepts, and Optimal Design

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

Abstract

There is growing theoretical and empirical support for the thought that we are at least sometimes warranted to infer to the best explanation of our evidence. This paper considers the question of whether we could ever be warranted in inferring to more than one best explanation, even though the best explanations are incompatible with each other. It is argued that a combination of Putnam's work on internal realism and recent insights from cognitive science on concepts, in particular on the notion of naturalness in relation to concepts, suggests a positive answer to this question.

Keywords: abduction; conceptual spaces; explanation; internal realism; natural kinds; optimal design.

There is a growing consensus that abduction is central to human reasoning (Douven & Schupbach, 2015; Schupbach, 2017; Williamson, 2018; Douven, 2019a, 2022). Roughly, abduction licenses us to infer that the best explanation of our evidence is true. There has been, and still is, much debate about how to make this rough idea precise. Here, we will focus on a question that so far has not been asked, to wit, whether we might ever have license to infer to more than one best explanation of our data, where these explanations are mutually exclusive. Naturally, we might have a perfectly good explanation of why Alice broke up with Bob in terms of how her feelings for him developed over time, while at the same time having a more scientific explanation in terms of Alice's personality traits, her childhood traumas, her past experiences with men. While very different, these explanations might strike us as being, each in its own way, entirely satisfactory. But note that these explanations could well co-exist. Our question does *not* concern this kind of situation. It concerns the kind of situation where there is more than one best explanation, and those explanations are *not* compatible. Could abduction warrant inferring any one of them?

The obvious answer would seem to be *no*, on grounds discussed in Lipton (1993) and Bird (2010). These authors hold that the best explanation must be significantly better than any available competitor before we can make the inference and accept the best explanation as true. This is a normative claim, but experimental research has shown that indeed people tend to infer to the best explanation only if it is clearly superior to the second-best

explanation available to them (Douven & Mirabile, 2018). And if two or more rival theories are tied for explanatory "bestness," then the aforementioned condition is arguably not satisfied so that we should refrain from making an abductive inference.

I want to explore the prospects of a positive answer to the question raised above. I will argue, tentatively, that there can be several mutually exclusive best explanations, and yet we may be licensed to infer any one of them. The answer to be proposed takes its cue from a remark that Quine (1992) makes in relation to the question of how to deal with situations in which theory choice is underdetermined not just by the currently available evidence but by all the evidence we might ever have. This kind of situation can arise when two or more theories are (what is called) empirically equivalent, which roughly means that they make the same predictions about the observable part of the world but make incompatible claims about what is going on behind the scenes.[1] I say "can arise," for two reasons, a boring one and a more interesting one. The boring reason is that the theories may be empirically inadequate—some of the predictions may be false—in which case the question of whether to infer any of them to be true is moot. The more interesting reason is that, at least from the perspective of a believer in abduction, of a number of incompatible theories making the same (correct) predictions, one may still offer a better explanation of the data than the others, which—from the said perspective—would warrant adopting the former at the expense of the latter. Note, however, the remaining possibility that we may encounter empirically equivalent theories that are also equally good explanations (Quine, 1975; Newton-Smith, 1981).

According to Quine's proposal, there is no need to choose among theories in this kind of situation. We are free to adopt any of them, albeit only one at any given time. In Quine's proposal, we are to conceive of the theories as different, equally legitimate, conceptualizations of reality, which may all be true (in a sense to be clarified). In practice, we may "oscillate" between these different conceptualizations "for the sake of added perspective from which to triangulate on problems" (Quine, 1992, p. 100). The idea is that each theory is "true in its conceptual scheme." While this remained only a suggestion in Quine's work, the idea is a cornerstone of Putnam's writings on internal realism from the 1980s and 1990s (e.g., Putnam, 1981, 1987, 1990). But even Putnam did not make much of an effort to clarify the notion of a conceptual scheme, nor did he do enough (in the eyes of critics) to alleviate the concern that truth-in-a-conceptual-scheme is a subjective notion that gives rise to an unpalatable form of relativism.

In the following, I aim to give content to the Quinean/Putnamian idea of there being alternative, yet equally valid conceptualizations of reality by drawing on the so-called conceptual spaces framework (Gärdenfors, 2000, 2014). In Decock and Douven (2012), it is shown how that framework can be used to render the notion of a conceptual scheme formally precise. While that paper mentioned concerns over internal realism amounting to relativism, it did not address those. Here, I will fill that gap by appealing to recent work on the *optimality* of concepts, notably, Douven and Gärdenfors' (2020) proposal that some conceptual schemes are better than others and that some are even optimal, where, however, the notion of optimality at play is that of Pareto optimality, meaning that there can be more than one optimal conceptual scheme (see also Douven, 2019b).

I start by summarizing Putnam's internal realism as well as the conceptual spaces

[1] For a precise statement of the problem of underdetermination, see Douven (2008).

framework and explain how the latter can be used to elucidate the former (Sect. 1). I then go into recent work on the optimality of conceptual spaces/schemes and explain how this work may help us arrive at a positive answer to the question of how we could ever be confronting two or more best explanations, where these explanations are mutually exclusive, and where it could be rational to infer any one of them (Sect. 2). Finally, I consider the question of whether the resulting position still leaves too much room for relativism (Sect. 3).

1 Internal realism and the conceptual spaces framework

1.1 Putnam's internal realism

Putnam's internal realism can be seen as an attempt to reconcile the realist intuition that the world is not of our making, that our believing things to be a certain way does in general not suffice to make them that way, with the antirealist thought that there are different yet equally valid ways of conceptualizing the world, and that which conceptual scheme (i.e., system of concepts) we use to think and talk about the world does contribute to "how things are." The proposed reconciliation is that, first, the conceptual scheme we use to think and talk about the world is not forced upon us by the world, and that how the world looks depends on the concepts in use. Putnam refers to this as "conceptual relativity," and in his view it should appeal to antirealists. But second—and this should appeal to realists—the world is the way it is, unaffected by what we believe about it, albeit that we must recognize that only from *within* a conceptual scheme can we make sense of the world being a certain way.

At the most fundamental level, internal realism is about whether the world has a "built-in structure," a structure determined by what the *natural properties* or *natural kinds* are; about, in other words, whether some things belong together in some metaphysically deep sense independently of whether we recognize them as belonging together. While Putnam does not dismiss the idea of there being natural properties, he does believe that all talk about such properties is only meaningful once a conceptual scheme is in place. In particular, he denies that the world itself singles out a conceptual scheme as being the one that *really* reflects the world's structure. To the contrary, in his view there can be "equally coherent but incompatible conceptual schemes which fit our experiential beliefs equally well," where none of these is privileged over the others (Putnam, 1981, p. 173).

Putnam (1987, p. 17 f) notes that "[c]onceptual relativity sounds like 'relativism'," but insists that it does not give rise to conceptual relativism or (as it is more commonly called) incommensurability, nor is tantamount to cultural relativism. Incommensurability does not follow because—Putnam claims——conceptual schemes can always be compared with one another, even if they are incompatible, and cultural relativism does not follow because not all conceptual schemes are on a par: there are better and worse ones.

What makes these claims hard to adjudicate is that Putnam has done little to clarify what exactly, in his view, a conceptual scheme is. He does say that it is "a way of speaking, a language" (Putnam, 1987, p. 36), but that is not particularly illuminating. It is no news that we can talk about the world in many different languages; surely that cannot be all there is to the idea of conceptual relativity. To maintain internal realism as a serious contender in the

realism debate, we need a precise answer to the question of what a conceptual scheme is, and this answer should imply that (i) different conceptual schemes can be incompatible with one another and yet at the same time be comparable, and (ii) not all conceptual schemes are equally good. Using the conceptual spaces framework, Decock and Douven (2012) propose an explication of the notion of conceptual scheme that meets these requirements.

1.2 The conceptual spaces framework

The guiding idea underlying the conceptual spaces framework is that concepts can be represented geometrically, as regions in similarity spaces. Similarity spaces are one- or multidimensional metrical spaces—sets of points on which a distance function or metric is defined—whose dimensions represent fundamental qualities in terms of which we may compare items with each other. Distances in such a space are meant to represent dissimilarities: the further apart the representations of two items are in the space, the more these items are dissimilar in whichever aspect the space is aimed to model.

While in principle any metric can be associated with a space, in practice only the Manhattan metric and the Euclidean metric are used. Both metrics are instances of the following schema, the former being the instance with $p = 1$, the latter the instance with $p = 2$:

$$\delta_S(x, y) \quad = \quad \sqrt[p]{\left(\sum_{i-1}^{n} |x_i - y_i|^p\right)}.$$

Here, S is an n-dimensional space and $x = \langle x_1, \ldots, x_n \rangle$ and $y = \langle y_1, \ldots, y_n \rangle$ are points in that space.

Most commonly, a similarity space is constructed on the basis of a number of pairwise similarity ratings (pairs of stimuli are shown to participants who are asked to indicate how similar those stimuli are), but confusion probabilities (data indicating how likely it is that two distinct stimuli are mistaken to be identical when flashed consecutively to a participant) and correlation coefficients (indicating how strongly answers to different questions "hang together") are also sometimes used. Such data are then first transformed into distances. In turn, these distances serve as input for a statistical dimension reduction technique, such as principal component analysis or, more commonly, multi-dimensional scaling (MDS), which output a space (Borg & Groenen, 2000; Hout, Papesh, & Goldinger, 2013; Abdi & Williams, 2010).

The aim is to obtain not just any spatial representation of the input data, but one that (i) is low-dimensional, preferably with no more than three dimensions; (ii) has dimensions we can make sense of by relating them to a fundamental attribute that the items used to generate the input data (e.g., the stimuli whose similarities were rated) can be said to instantiate to different degrees, where preferably the "fundamentality" of he attribute can be explained by reference to certain properties of our perceptual or cognitive apparatus; and (iii) has good model fit, basically meaning that it provides an accurate representation of the input data (e.g., if the input data were similarity judgments, then the more similar two items are, the closer should the points representing them be in the output space). While we will not always be able to obtain a space satisfying these criteria, by now there are a great number of similarity spaces to be found in the literature that do check all the boxes.

To be clear, similarity spaces are *not* conceptual spaces: they represent similarities, not concepts. Rather, conceptual spaces are built on top of similarity spaces. There are different ideas about how to get a conceptual space from a similarity space, but the approach that has come to dominate the field turns similarity spaces into conceptual spaces by deploying a combination of prototype theory and the mathematical technique of Voronoi tessellations (Gärdenfors, 2000, 2014). Central to prototype theory is the thought that instances of a concept can be representative of it to differing degrees, with the most representative one being the concept's prototype (Rosch, 1973, 2011). And given a space and a set of designated points in that space, we can create a Voronoi tessellation on the space by dividing it into disjoint cells such that each cell is associated with precisely one of the designated points and contains those and only those points in the space that are at least as close to that designated point as they are to any of the other designated points (for details, see Okabe et al., 2000). The recipe for turning a similarity space into a conceptual space is now simply this: Identify the points in the space that are prototypical of the concepts the space is supposed to represent and use these as the designated points for producing a Voronoi tessellation of the space. Each of the cells represents a concept.

To illustrate, CIELab space and CIELuv space are widely used as color similarity spaces.[2] Both are spindle-like three-dimensional spaces, with one dimension—the vertical axis—representing luminance (or brightness), which goes from white to black through various shades of gray; the second dimension being what is commonly known as "the color wheel," which goes through blue, violet, red, orange, yellow, and green, to arrive at blue again, with each color gradually blending into the next; and the third dimension being saturation, which indicates how intense or deep a shade is. To make a conceptual color space out of either similarity space, one can locate the various prototypical colors in CIELab/CIELuv space, and then use those to define a Voronoi tessellation on that space. This allows us to think of the concept RED as a region in CIELab/CIELuv space, to wit, the region inside the cell associated with the RED prototype.[3]

As a disclaimer, I note that it is still unknown what exactly the scope of the conceptual spaces approach is. So far, most applications have been to families of perceptual concepts.[4] However, there is also some work on representing more abstract concepts in conceptual spaces, such as Gärdenfors' (2007) work on action concepts, Gärdenfors and Warglien's (2012) work on event concepts, and Oddie's (2005) and Verheyen and Peterson's (2021) work on moral concepts. There is even some work on still more abstract, scientific concepts like mass and acceleration; see Gärdenfors and Zenker (2011, 2013). Nevertheless, at this point, it is prudent to be cautious and not oversell the conceptual spaces approach. It is a

[2] Which of the two is used depends on the viewing conditions. The former works better for colors on paper or cloth, while the latter gives better results when the colors are shown on screen.

[3] In the standard conceptual spaces framework, as found in Gärdenfors (2000, 2014), concepts are *well-delineated* regions in similarity spaces. It is readily appreciated, however, that that can hold only by way of idealization, at least as a general claim. For instance, color concepts tend to be vague, in that there are shades which neither entirely fall under a concept nor entirely do *not* fall under it. See Douven et al. (2013) for how to extend the conceptual spaces framework so that it can accommodate vagueness. For empirical research supporting the descriptive adequacy of the extension, see Douven (2016, 2019c, 2021), Douven et al. (2017), and Verheyen and Égré (2018).

[4] For instance, see Petitot (1989) for relevant work on auditory concepts; Castro, Ramanathan, and Chennubhotla (2013) for work on olfactory concepts; and Gärdenfors (2000), Churchland (2012), and Douven (2016a, 2021) for work on shape concepts.

real possibility that the conceptual spaces approach is only going to be part of the story about concepts and that a "final" theory of concepts is going to be hybrid and only partly similarity-based (other parts might, for instance, be rule-based; see Hahn & Chater, 1997, 1998). Philosophers have a penchant for general theories. While I see the attraction of such theories, I believe that the said penchant often stands in the way of making progress. For instance, in Douven (2016a) I argued that one reason why many semantics of conditionals have fared so badly, in terms of both broad acceptance and empirical validation, is that they are meant to apply to each and every way in which the word "if" is used in our language. Similarly, in Douven (1998, 2016b) I argued against semantics that try to explain sentence meaning in terms of one key concept (usually either truth or verification), without being open to the possibility that we need a different semantics for different parts of our language; so, for instance, a different semantics for the language of mathematics, or physics, than we need for the more broadly shared parts of our language. I likened the preference for a uniform semantics to the preference for an explanation of every disease in terms of at most a few fundamental concepts. If simplicity and elegance were what mattered most in scientific theories, such a uniform theory of disease would win hands down from the hodge-podge of local explanations that are now to be found in the medical literature. Yet no one believes that we would be better off with the highly uniform theory. For all we know, there is no simple and elegant, uniform theory of diseases that is also helpful in any way. Similarly, for all we know, there is no simple and elegant, uniform theory of concepts that is worth having.

1.3 From conceptual spaces to conceptual schemes

Decock and Douven (2012) propose to use the conceptual spaces framework to elucidate the notion of a conceptual scheme and thereby to place internal realism on a more solid footing. Concretely, they propose to identify a conceptual scheme with a set of conceptual spaces. Thus, in their proposal a given conceptual scheme could, for instance, consist of a color space, an auditory space, several shape spaces, and many more besides, where each of those spaces has an associated set of prototypes that determine which concepts are being represented in the space.

As Decock and Douven note, their proposal has a number of attractive features. For instance, it turns Putnam's thesis of conceptual relativity into a precise statement with empirical content. And with regard to Putnam's claim that there is no one best conceptual scheme, Decock and Douven note that, in their proposal, (i) conceptual schemes can differ from each other in the type and number of conceptual spaces that they contain as well as in the geometry and topology of those spaces, and (ii) there is a wealth of empirical evidence that conceptual spaces in actual use *do* differ, not only between cultures, but also at an individual level among members of the same culture.[5]

Another advantage of the proposal is that it now becomes easy to see how different conceptual schemes can be incompatible with each other. Suppose two conceptual schemes both contain a color space, where however these differ in their topological structure, perhaps because the color spaces are associated with different sets of prototypes. Then one space could classify a particular color shade as, say, definitely blue which the other classifies

[5] For some particularly compelling evidence, see Regier, Kay, and Khetarpal (2007) and Douven et al. (2022).

as definitely green. In that case, the schemes would give rise to incompatible verdicts about the shade.

Decock and Douven further point out that, in their proposal, Putnam can easily be seen to be right in claiming that conceptual relativity amounts to neither conceptual relativism nor cultural relativism. As regards the former, they note that the conceptual spaces framework offers a kind of meta-perspective from which one can compare conceptual schemes, for instance in terms of shared and non-shared conceptual spaces. As for cultural relativism—the claim that one conceptual scheme is as good as another—it is not difficult to think of sets of conceptual spaces that are too poor to serve our purposes (e.g., because they leave out some crucial conceptual spaces) or which include spaces whose topology hinders rather than helps the learning or memorization of concepts.

Nothing found in the literature on conceptual spaces excludes the possibility of there being more than one best conceptual scheme, which may be enough for many realists to keep objecting to internal realism, Decock and Douven's precisification notwithstanding. Indeed, I expect that realists will want to hold that, whichever conceptual schemes people may use, there is but one that captures the true nature of reality. Specifically, many realists will insist that there is one set of conceptual spaces that we *should* all use if our aim is to represent reality as it is—the set consisting of those spaces representing concepts that match the *natural kinds* in the world.

Only recently have researchers working on conceptual spaces delved into the question of what makes a concept a natural one. This interest has led to an account of naturalness that accommodates realist intuitions, at least to some extent. This account makes central the notion of an optimally designed conceptual space.

2 The optimal design theory of natural concepts

We saw that, in the conceptual spaces approach, concepts are regions in similarity spaces. In principle, any region in a similarity space can represent a concept. But not any region in a similarity space is a candidate for representing a concept that we might ever have a use for. Indeed, pick any region in a similarity space, and almost certainly it will fail to correspond to a concept that has ever figured, or will ever figure, in our thinking. As Gärdenfors (2000) pointed out early on, we are only interested in *natural* concepts.

However, at the time, Gärdenfors was not prepared to commit to any definition of naturalness and offered only what he saw as a necessary but insufficient condition for a region to be natural, to wit, convexity, which is satisfied by a region if and only if, for any pair of points in it, every point lying between those points lies in the region as well. Gärdenfors presents the convexity requirement as a principle of cognitive economy. Given the memory and processing limitations humans are subject to, it is much easier for us to deal with convex regions than with arbitrarily shaped ones. He also cites empirical evidence supporting the requirement: concepts in actual use do tend to correspond to convex regions in the relevant similarity spaces.

Gärdenfors' preferred way of obtaining a conceptual space from a similarity space is the one described in Section 1.2: locate the prototypes in the similarity space, and then apply the technique of Voronoi tessellations to carve up the space into regions, which are then said to represent the concepts. This has the nice side effect of guaranteeing convexity, given

that, as a matter of mathematical fact, all cells in a Voronoi tessellation are convex (Okabe et al., 2000). By the same token, however, we can also easily appreciate why convexity is not even *close* to being sufficient for naturalness, for the mathematical result holds given *any* set of points in the space that we might use to generate a Voronoi tessellation. For instance, take some random set of points in CIELab space, use these to tessellate the space, and you will end up with a set of convex regions in color space. Most likely, those regions will appear gerrymandered to us, and we will be unable to recognize them as representing natural color concepts.

The question of which conditions to add to convexity to arrive at a characterization of natural concepts was taken up in Douven and Gärdenfors (2020). These authors took their cue from design thinking in modern biology, which explains biological processes in organisms or biological traits in terms of good engineering design, the idea being that such processes and traits are exactly as one would expect them to be if they had been designed by a team of good engineers (e.g., Alon, 2003; Nowak, 2006). In a nutshell, Douven and Gärdenfors' proposal is that this idea of good design also makes sense in relation to conceptual spaces, and that a natural concept is one that is represented by a cell of an optimally designed conceptual space.

Already the convexity criterion is plausibly thought of as a design principle: If one were tasked with designing a conceptual architecture for a similarity space, one would want it to yield convex concepts, for the reasons of cognitive economy mentioned above. Douven and Gärdenfors state a number of additional similarly motivated design principles. Jointly, these principles amount to two broad requirements, to wit, that a space should have the right granularity, and that it should allow for having prototypes that are both good representants and easily distinguishable.

The granularity requirement means that a space should not be partitioned too finely in order to avoid overtaxing the user's memory, but at the same time should be partitioned finely enough to allow the user to make and communicate sufficiently many distinctions. Also, we should find this balanced granularity throughout a space: in general, it should not be the case that we can make very fine-grained distinctions in one part of a similarity space but then only rather coarse-grained ones in other parts.

The requirement concerning prototypes is that, on the one hand, we should be able to spread the prototypes out in the space, so that the user will not be tempted to mistake one for another, while on the other hand, we should be able to place the prototypes such that each is a good representant of all the other items falling within the concept of which it is the prototype. In short, the prototypes should be as dissimilar to each other as is allowed by the geometry of the similarity space, but they should also be as similar as is possible to each of the items they are supposed to exemplify.

The foregoing is a rather abstract summary of Douven and Gärdenfors' proposal. To see more concretely what it amounts to, here is a first illustration, using Liljencrants and Lindblom's (1972) research on vowel systems, which Douven and Gärdenfors cite as an important source of inspiration for their proposal. Liljencrants and Lindblom start from the observation that, although the human vocal tract is, in principle, capable of producing indefinitely many different vowels, study of the vowels found in spoken languages reveals that only a handful of those are actually instantiated. Why is that?

Vowels can be represented in a three-dimensional similarity space. Liljencrants and

Lindblom use this space to tackle the foregoing question. More exactly, their hypothesis is that we tend to find the same vowels across languages because those vowels maximize contrast. The hypothesis makes prima facie sense because by optimizing contrast among vowels, we minimize the risk of mistaking one vowel for another and thereby minimize the risk of miscommunication. In terms of optimal design: the hypothesis is that the constellation of locations in vowel space that instantiate actually used vowels is one which clever engineers would have picked as well.

Liljencrants and Lindblom went on to test their hypothesis via computer simulations. They wrote a computer program to calculate for a given number n the constellation of n points in vowel space that maximizes, for that number of points, the total distance among the points and so maximizes the contrast among the vowels represented by those points. They then looked at languages with numbers of vowels varying from three to twelve and compared their computational results with the constellations of points in vowel space corresponding to the vowels found in the various languages. For languages with up to six vowels, the results were extremely accurate. For languages with more vowels, there were more errors. Liljencrants and Lindblom explain this fact by noting that their computer simulations look only at contrast among vowels while in reality other factors may also play a role in the selection of vowels. They in particular mention the possibility of articulatory factors being involved as well: "a vowel system which has been optimized with respect to communicative efficiency consists of vowels that are not only 'easy to hear' but also 'easy to say' " (Liljencrants & Lindblom, 1972, p. 856).

Another illustration is to be found in Douven (2019c), which explicitly sought to empirically test Douven and Gärdenfors' proposal. This work focused specifically on the part of the proposal according to which an optimally partitioned similarity space allows the placement of prototypes that are both highly representative of the other items in their concept and easy to distinguish from the other prototypes in the space, in order to minimize the chance that users make classification errors. The experiment reported in Douven (2019c) relied on color similarity space and on knowledge of the partitioning of that space into the eleven concepts corresponding to the so-called Basic Color Terms (Berlin & Kay, 1969) that was documented in Jraissati and Douven (2018).

The experiment aimed to answer the question of whether the constellation of basic color prototypes satisfies the design principles of good representativeness and good discriminability. To that end, participants were asked to identify the shades that, in their opinion, were typical for red, green, blue, and so on. In a next step, the responses per basic color were "averaged" (by taking the center of mass of their coordinates in color space), and those averages were taken as good indicators of the locations of the basic color prototypes. These results were compared with 5,000,000 randomly generated constellations of potential prototypes of the eleven basic colors and it was found that, in over 99.99 percent of those constellations, whenever they did better on the count of representativeness, they did worse on the count of contrastiveness, suggesting that the actual constellation was a (near to) Pareto optimal trade-off of those two desiderata. In a further step, the actual constellation of prototypes was also compared with the outcomes of a computational procedure somewhat similar to the one Liljencrants and Lindblom had used, although they had only sought to maximize contrastiveness among the vowels, while the procedure described in Douven (2019c) aimed to find the best trade-off between contrastiveness among the basic

color prototypes and representativeness of those same prototypes. This, too, yielded strong evidence that the actual constellation is Pareto optimal.

3 Natural kinds, really?

We started with the question of whether we could ever have two best explanations of the available evidence, where these explanations are incompatible and yet we can warrantedly infer either one of them, or in fact even both, although only individually at different times. A remark in Quine's work, and more substantively Putnam's work on internal realism, suggested a positive answer to that question. The challenge was to make that answer look *attractive*.

The reformulation of internal realism using the conceptual spaces framework offered in Decock and Douven (2012), and briefly recapped in Section 1.3, was meant to at least alleviate concerns about whether the notion of a conceptual scheme, which is key to internal realism, can be given a rigorous formulation. We saw that conceptual schemes can be understood as collections of conceptual spaces, which are well-defined mathematical entities. But at the end of Section 1.3, we also mentioned the concern that internal realism might be unable to account for a thought that not only characterizes traditional realism but also strikes many as utterly commonsensical, to wit, that there is a *right* conceptual scheme—the one whose concepts correspond to natural kinds—and that that is the one we should use for talking and theorizing about the world.

In the previous section, I have summarized the optimal design account of natural concepts because I believe this will help us address the concern about natural kinds. Unsurprisingly, my suggestion is that natural kinds are the worldly correlates of natural concepts, understood in the manner of the optimal design account. But how plausible is this? In standard realist thinking, there could never be more than one conceptual scheme capturing the natural kinds. And it is not clear that the optimal design account guarantees satisfaction of this uniqueness condition. In fact, if it did, then what would remain of the Quinean–Putnamian idea that we can oscillate between equally valid descriptions of reality, which we seek to make look plausible in this paper?

We mentioned that the empirical results reported in Douven (2019c) established that the actual constellation of color prototypes is Pareto optimal. That means one cannot find a constellation that does better both in terms of how contrastive the prototypes are (i.e., how dissimilar the prototypes are to each other) and in terms of how representative they are (i.e., how similar the prototypes are to the items they are meant to represent). However, there do exist constellations that cannot be said to make *worse* trade-offs between contrastiveness and representativeness than the actual constellation does. Some do a bit better than the actual constellation with respect to contrastiveness; others do a bit better with respect to representativeness. These alternative constellations are thereby also Pareto optimal.

If contrastiveness and representativeness do not fix a unique constellation of color prototypes, and so a fortiori do not fix a unique conceptual space for color concepts, then perhaps together with some or all of the other design principles proposed in Douven and Gärdenfors (2020) they *do*. Perhaps, though I am not hopeful in this regard. The reason is that there will only be *more* trade-offs to be made. It is not just that contrastiveness and representativeness can pull in different directions; the principles concerning the

granularity of the partitioning of color space pull in different directions by definition. For instance, we would like to be able to express very fine-grained distinctions among colors—and thus have many color concepts—but we also want to avoid putting too much strain on memory, and so try to get by with relatively few color concepts.

It thus appears that, most fundamentally, the challenge for the present proposal is to clarify how the optimal design account's notion of natural concepts can be rightfully said to reflect the structure of reality. Realists and nominalists have been debating the nature of what we call "natural kinds" for ages, the former maintaining that natural kinds are classes of things that *objectively* belong together because they carve nature at its joints, and the latter objecting that, for all anyone has ever shown, nature is jointless, and that we should feel free to carve where we want; what *appear* to be nature's joints are really divisions of our own making.

The realists always seemed to have the upper hand precisely because, well, natural kinds do appear natural to us. What could be more natural than how we group colored things, animals, metals, and so on, into different categories? Still, a major problem for realists is to explain how nature could do so much as privilege certain classes over others. It was long believed that modern science would be able to provide the requisite explanation, by discovering the micro-essence of each natural kind—appealing to shared DNA, or molecular structure, or atomic number, or what have you—and that this micro-essence would account for the kind's phenomenal properties which made us consider it to be *natural*. But this project did not go quite as expected. The micro-essentialist answer that science appeared to give proved contentious under scrutiny. For instance, while we regard cows to constitute a natural kind, the bovine genome is not fixed once and for all but is subject to changes, due to evolutionary pressures (Ghiselin, 1987; Dupré, 1993; Sterelny & Griffiths 1999). And the claim that water is H_2O is a gross simplification; in reality, water is a mixture of H_2O, D_2O, and a number of other isotope combinations of hydrogen and oxygen (van Brakel 1986, 2005; Needham 2000, 2011; Weisberg 2005). Such considerations led Churchland (1985, 12 f) to conclude that natural kind concepts are much sparser than had been generally believed and only concern fundamental physical entities and quantities, like neutrons, quarks, charge, mass, and momentum.

But adopting such a minimalist stance vis-á-vis natural kind concepts robs realism of much of what had made it intuitively appealing. Indeed, biological and chemical kinds serve as the primary examples of natural kinds in Putnam (1975) and Kripke (1980), two publications pivotal in rekindling twentieth-century philosophers' interest in the realism debate. And color concepts figure prominently as examples of (what she calls) natural categories in Rosch (1973), which has been highly influential in psychology.

On the other hand, Leslie (2013) musters a vast amount of evidence from developmental psychology indicating that our essentialist intuitions may well be due to inchoate cognitive biases and may thus "reflect only facts about us, not facts about the deep nature of reality" (p. 158). Perhaps we simply have to get over the failure of the micro-essentialist program and learn to live with something like Churchland's minimalism.

However, contemplation on the role natural kind concepts play in science may stir more serious concerns about Churchland's position. A metaphysical idea that guides science and that, according to many, is at the same time buttressed by the instrumental success of science, is that of a world hierarchically organized, where the different levels

of organization are not only internally structured—into biological kinds, chemical kinds, physical kinds, and so on—but also interlock in systematic ways, via causal, functional, and part–whole relationships. Darden and Maull (1977) point out the vital importance of these interrelations for the practice and, ultimately, the success of science (see, in the same vein, Shapere in Callebaut, 1993, p. 159 ff). The role these interrelations play in science would be difficult to make sense of if we were realists about physical kinds, perhaps, but then were to side with the nominalists on biological and chemical kinds and hold that these are mere arbitrary groupings.

In the present proposal—basically, internal realism cashed out within the conceptual spaces framework, and then with an optimal design twist added to it—natural kinds are said to be nonarbitrarily grouped classes, without however conceding to the realist that there is necessarily a unique best description of the world, one which depicts the world as seen from a God's eye viewpoint (to use one of Putnam's favorite phrases). Natural kinds are nonarbitrary because not every way of dividing up the world is optimal, from an engineering perspective. Indeed, almost all partitionings of a similarity space will result in a non-optimal conceptual space, meaning that, even though not *unique*, natural kinds should still be *sparse*.

Still, have we not sacrificed the idea that there is an *objective* world out there, independent of our conception of it? I think not. "The mind and the world jointly make up the mind and the world," so goes the slogan that Putnam (1981, p. xi) famously used to summarize internal realism. As intended by Putnam, the word "mind" in the slogan refers to our mental activities, which in his view contribute to what the world looks like. The slogan could also be used to summarize the position advanced in the present paper, although then "mind" is to be taken to refer to the constraints under which the human mind has to operate, to what must be the case for our mental activities to operate in the best possible manner, where various limitations our mental apparatus is subject to, in conjunction with the pressures we face in our perpetual struggle for existence, determine what is "best possible."

More specifically, in claiming that natural concepts are those that populate an optimally designed cognitive system, we understand "optimality" as being defined by reference to broad constraints we humans labor under. Douven and Gärdenfors (2020) argue that our conceptual systems should facilitate learning and memorization, and also help to avoid classification errors, and moreover do all of this in a cost-effective manner. Thereby, they make reference to our limitations: had our memories unlimited storage capacities, or were our discriminatory capacities much greater than they are in reality, there might be much less concern about the architecture of our conceptual systems—we might get by on many such systems, no matter the details of their design, and cost considerations might be much less pressing.

This proposal manifestly makes natural concepts relative to us humans. However, it does not make natural concepts relative to any specific culture, or to any transient interests we may have, or to whichever context we may happen to speak or theorize within. There should thus be no concern about our position being relativist in any of the potentially damning senses that Putnam's is, according to some critics (see, e.g., Devitt, 1991, Ch. 12).

To the contrary, conceptual systems can lay claim to objectivity inasmuch as we come to choose neither the similarity spaces nor the constraints under which our mind is to operate, and which motivate the design principles proposed in Douven and Gärdenfors (2020); we

had, and have, no say over the make-up and functioning of our perceptual and cognitive apparatuses. The current proposal could not be further removed from Goodmanian ideas of worldmaking (Goodman, 1978) and similar approaches to metaphysics which leave a lot of room for decision making.

To be sure, "objective," on our proposal, does not imply eternal or otherwise immutable: the same pressures that have shaped our conceptual systems may also reshape them, for instance, because some similarity spaces may change (e.g., our perceptual apparatus may change), or the constraints the mind is under may change. But no such strong sense of objectivity may be needed to make sense of the role concepts play in our thinking, not even in science. Science is our best attempt to make sense of the world—sense for us, from a human perspective, not from a God's eye viewpoint. This is a task we tackle, and cannot but tackle, using our concepts, and the view of concepts taken on board in this paper makes it entirely possible for us to claim that there is a best set of concepts for this task, even if that set may not be unique. Science is then still an endeavor in which we try to figure out which systematic relations hold among the various natural concepts. The end result, if we succeed in this endeavor, will have a claim to objectivity, even if not in the grandiose, Platonic sense traditionally envisioned by realists. But Plato's heaven may have been a philosophical fiction all along. Realists who are nonetheless dissatisfied with our proposal should ask themselves what surplus explanatory work a Platonic notion of objectivity could do. I am unable to think of any. (If the answer is that such a notion would better explain your intuitions, ask yourself why nature should care about those.)

I end this section by mentioning two reasons why realists should actually *like* the optimal design take on natural kinds. First, realists have appealed to natural kinds in trying to block Putnam's (1980) model-theoretic argument against realism. In a nutshell, the argument purports to show that, for realists, truth amounts to no more than consistency. By some well-known results from model theory, any consistent theory has a model and, given some plausible assumptions, it has a model whose domain contains as many objects as there are in the world. The core of the argument is that the realist is in no position to reject a one-to-one mapping from a theory's model onto the world as being unintended. Realists have objected to the argument that there is no guarantee that the one-to-one mapping that the argument shows to exist also gets the world's *structure* right, where this means that the extensions of the theory's predicates assigned by the model map onto natural kinds (Merrill, 1980; Lewis, 1983).[6] To which Putnam retorted that the idea of a built-in structure, of there being natural kinds independently from human thinking and theorizing, makes no sense; it is—in his view—only from within a conceptual scheme that the notion of natural kinds can be understood. In trying to argue to the contrary, realists face the problems mentioned above. The optimal design proposal can help out at this point. In this proposal, natural kinds are still sparse, as said, and so there is no guarantee that the mapping Putnam constructs in his model-theoretic argument maps the predicates of the language onto natural kinds. At the same time, the proposal gives content to the notion of natural kinds without invoking micro-essences, while still leaving the idea of natural kinds being objective intact (in the sense of "objective" explained above).

A second advantage of the optimal design proposal is that it provides a straightforward

[6]This was not the only response to Putnam's argument. See Devitt (1991, Ch. 12), Douven (1999a, 1999b), and Button (2013) for discussion.

response to an argument that is meant to favor nominalism over realism and that is to be found in Book III of Locke's *An Essay Concerning Human Understanding* from 1689. There, Locke propounds an empirical argument for nominalism, based on the best science available at the time. He addresses the rarely asked question of what constitutes the "joints of nature" which, according to Plato, separate the various natural kinds from one another. Locke's answer is that, if they exist, there have to be "Chasms, or Gaps" (III, vi, 12) between different classes of entities; these would separate the various classes, thereby structuring the world in an objective fashion. But, Locke argues, when we look at the world, we see that the requisite gaps are just not there. Wherever we suspect one, we see that there are intermediate cases, closing the gap, so to speak, to find, ultimately, that things "differ but in almost insensible degrees" (*ibid*.).

But consider again the case of color, which provides us with an uncontentious example of a gapless domain. It does not require sophisticated software to have your computer screen show a patch that is clearly green (say) and then have its color change seamlessly to clearly blue, or clearly yellow, or whichever color you prefer. Still, the fact that this domain is continuous does not render the optimal design account inapplicable. In fact, color space serves as one of the main examples in Douven and Gärdenfors (2020). What this means is that, at least in the color domain, the joints of nature are constituted by the shape of the relevant similarity space—a shape which depends on the human perceptual apparatus—in conjunction with various principles of optimal design, which depend on our cognitive makeup. Jointly, similarity and optimization thereby fix, nonarbitrarily, the structure of the color domain, even if, as explained above, that structure has no place in Platonic metaphysics.[7]

4 Conclusion

We asked whether we could ever be in a position where we are warranted in inferring more than one best explanation, where the best explanations are incompatible. We explored the prospects for a positive answer, building on Putnam's work on internal realism. While little enthusiasm for that work can be found in today's philosophical literature, I hope to have shown that, at a minimum, it deserves another chance. As already shown in Decock and Douven (2012), the conceptual spaces framework can help greatly to make mathematically precise Putnam's rather loosely stated thoughts on conceptual schemes. But Decock and Douven did not address the concern that internal realism might amount to a form of relativism that would seem incompatible with our intuitions about natural kinds (e.g., that they are objective, and that they are robust enough to play a central role in modern science). I have argued that, at this point, the optimal design account of natural kinds can come to the rescue. According to this account, natural kinds are the worldly correlates of natural concepts, where the latter are those concepts that are represented by optimally designed conceptual spaces. What counts as optimal design is relative to our perceptual apparatus as well as our cognitive makeup, but inasmuch as neither is up to us, it is not up to us either

[7] To keep things simple, I have skipped the issue of how to represent vagueness within the conceptual spaces framework. For how this can be done, see the papers cited in note 3. Results reported in those papers suggest an explanation of Locke's intuition that there are gaps among kinds in terms of boundary regions in conceptual spaces. Again, I leave this aside for now.

what the natural kinds are.

The optimal design account of natural kinds is perfectly compatible with there being more than one optimal conceptual scheme. Indeed, I would be surprised if design principles were able to fix a uniquely best color space, a uniquely best taste space, a uniquely best olfactory space, and so on. Admittedly, however, I cannot entirely exclude that they can do that after all. So it is only with some caution that I side with Quine and Putnam in thinking that we can be faced with mutually exclusive theories which appear equally good explanations and we can rationally adopt either, or any one, of them.

But supposing we can be faced with such theories, could we ever be warranted in *simultaneously* adopting two or more of them as best explanations?[8] It depends on what we mean by "adopt." If it means recognizing both (or all) theories as being equally adequate, empirically and theoretically, as building on different conceptual systems which, however, are both (or all) Pareto optimal, then the answer is positive, as far as I can see. We can think of both (or all) theories as telling us the truth about the world, or about a certain part of the world, while requiring us to activate different yet equally natural concepts. If, on the other hand, by "adopt" we mean using both (or all) theories simultaneously as a basis for further research, for developing new theories, for designing experiments, and so on, then the answer is less clear to me. For philosophers, it is easy to write about theories in the abstract and to recommend how scientists should go about testing their theories and especially about how scientists ought to decide which theories to accept. In scientific practice, however, it can take a lot of time and effort to familiarize oneself enough with even just one theory to feel comfortable working out its empirical consequences and conceiving experiments aimed to test those consequences. As a result, it is rare to see a scientific paper presenting evidence meant to discriminate among more than two or three rival theories.[9] That practice is understandable and even justified, for the reasons mentioned. The situation is not very different with regard to conceptual schemes. One may be willing to admit that other ways of carving up (say) color space than the one we have gotten used to are equally optimal and therefore could lay as much claim to being "natural" as the familiar one. But precisely because the way we commonly carve up color space is the one we are familiar with, it may not make a lot of sense, and may actually be counterproductive, to ever adopt any other system of color concepts. It would thus seem reasonable to use the theories we are familiar with, which build on a conceptual system we feel at home in, as a framework for conducting further work, even if there are alternatives that we must acknowledge as providing equally good explanations of our evidence.[10]

References

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *WIREs Computational*

[8] Thanks to Jonah Schupbach for raising this question.

[9] For instance, in the area of science that I know best—the psychology of conditional reasoning—I have *never* seen a paper in which an account of conditionals is compared with *all* its extant rivals. Typically, the theory in which the authors have a stake is compared with two, at most three, of what according to the authors are its most serious contenders (e.g., the suppositional theory is compared with the mental models account and with certain versions of inferentialism, leaving many of the known theories of conditionals undiscussed).

[10] I am greatly indebted to David Glass and Jonah Schupbach for valuable comments on a previous version of this chapter.

Statistics 2: 433–459.

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

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

Bird, A. (2010). Eliminative abduction: Examples from medicine. *Studies in the History and Philosophy of Science* 41: 345–352.

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

Button, T. (2013). *The Limits of Realism*. Oxford: Oxford University Press.

Callebaut, W. (1993). *Taking the Naturalistic Turn*. Chicago: Chicago University 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, doi: 10.1371/journal.pone.0073289.

Churchland, P. M. (1985). Conceptual progress and word/world relations: In search of the essence of natural kinds. *Canadian Journal of Philosophy* 15: 1–17.

Darden, L. & Maull, N. (1977). Interfield theories. *Philosophy of Science* 44: 43–64.

Decock, L. & Douven, I. (2012). Putnam's internal realism: A radical restatement. *Topoi* 31: 111–120.

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

Devitt, M. (1991). *Realism and Truth*. Oxford: Blackwell.

Douven, I. (1998). Truly empiricist semantics. *Dialectica* 52: 127–151.

Douven, I. (1999a). Putnam's model-theoretic argument reconstructed. *Journal of Philosophy* 96: 479–490.

Douven, I. (1999b). A note on global descriptivism and Putnam's model-theoretic argument. *Australasian Journal of Philosophy* 77: 342–348.

Douven, I. (2008). Underdetermination. In S. Psillos & M. Curd (eds.) *The Routledge Companion to Philosophy of Science* (pp. 292–301). London: Routledge.

Douven, I. (2016a). *The Epistemology of Indicative Conditionals*. Cambridge: Cambridge University Press.

Douven, I. (2016b). Rethinking Semantic Naturalism. In S. Goldberg (ed.) *The Brain in a Vat* (pp. 174–189). Cambridge: Cambridge University Press.

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

Douven, I. (2017). How to account for the oddness of missing-link conditionals. *Synthese* 194: 1541–1554.

Douven, I. (2019a). Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence* 275: 235–251.

Douven, I. (2019b). The rationality of vagueness. In R. Dietz (ed.), *Vagueness and Rationality* (pp. 115–134). New York: Springer.

Douven, I. (2019c). Putting prototypes in place. *Cognition* 193: 104007, doi: 10.1016/j.cognition.2019.104007.

Douven, I. (2021). Fuzzy concept combination. *Fuzzy Sets and Systems* 407: 27–49.

Douven, I. (2022). *The Art of Abduction*. Cambridge MA: MIT Press.

Douven, I. & Decock, L. (2017). What verities may be. *Mind* 126: 386–428.

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

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

Douven, I. & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Language, Memory, and Cognition* 44: 1792–1813.

Douven, I. & Schupbach, J. N. (2015). The role of explanatory considerations in updating. *Cognition* 142: 299–311.

Douven, I., Verheyen, S., Elqayam, S., Gärdenfors, P., & Ost-Vélez, M. (2022). Similarity-based reasoning in conceptual spaces. Manuscript.

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

Dupré, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge MA: Harvard University Press.

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

Gärdenfors, P. (2007). Representing actions and functional properties in conceptual spaces. In T. Ziemke, J. Zlatev, & R. M. Frank (eds.) *Body, Language and Mind* (Vol. 1, pp. 167–195). Berlin: De Gruyter.

Gärdenfors, P. (2014). *The Geometry of Meaning*. Cambridge MA: MIT Press.

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

Gärdenfors, P. & Zenker, F. (2011) Using conceptual spaces to model the dynamics of empirical theories. In E. J. Olsson & S. Enqvist (eds.) *Belief Revision Meets Philosophy of Science* (pp. 137–153). New York: Springer.

Gärdenfors, P. & Zenker, F. (2013) Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. *Synthese* 190: 1039–1058.

Ghiselin, M. (1987). Species concepts, individuality, and objectivity. *Biology and Philosophy* 2: 127–143.

Goodman, N. (1978). *Ways of Worldmaking*. Harvester Press.

Hahn, U. & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (eds.) *Knowledge, Concepts, and Categories* (pp. 43–92). Hove UK: Psychology Press.

Hahn, U. & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Distinguishable? *Cognition* 65: 197–230.

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *WIREs Cognitive Science* 4: 93–103.

Jraissati, Y. & Douven, I. (2017). Does optimal partitioning of color space account for universal categorization?' *PLoS ONE* 12: e0178083, https://doi.org/10.1371/journal.pone.0178083.

Kripke, S. (1980). *Naming and Necessity*. Oxford: Blackwell.

Leslie, S.-J. (2013). Essence and natural kinds: When science meets preschooler intuition. *Oxford Studies in Epistemology* 4: 108–166.

Lewis, D. K. (1983). New work for a theory of universals. *Australasian Journal of Philosophy* 61: 343–377.

Liljencrants, J. & Lindblom, B. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48: 839–862.

Lipton, P. (1993). Is the best good enough? *Proceedings of the Aristotelian Society* 93: 89–104.

Lipton, P. (2004). *Inference to the Best Explanation*. London: Routledge.
Merrill, G. H. (1980). The model-theoretic argument against realism. *Philosophy of Science* 47: 69–81.
Needham, P. (2000). What is water? *Analysis* 60: 13–21.
Needham, P. (2011). Microessentialism: What is the argument? *Noûs* 45: 1–21.
Newton-Smith, W. H. (1981). *The Rationality of Science*. London: Routledge.
Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge MA: Harvard University Press.
Oddie, G. (2005). *Value, Reality, and Desire*. Oxford: Oxford University Press.
Okabe, A., Boots., B., Sugihara, K., & Chiu, S. N. (2000). *Spatial Tessellations* (2nd ed.). New York: Wiley.
Petitot, J. (1989). Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics* 15: 25–71.
Putnam, H. (1975). The meaning of "meaning." *Minnesota Studies in the Philosophy of Science* 7: 131–193.
Putnam, H. (1980). Models and reality. *Journal of Symbolic Logic* 45: 464–482.
Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.
Putnam, H. (1987). *The Many Faces of Realism*. La Salle IL: Open Court.
Putnam, H. (1990). *Realism with a Human Face*. Cambridge MA: Harvard University Press.
Quine, W. V. O. (1975). On empirically equivalent systems of the world. *Erkenntnis* 9: 313–328.
Quine, W. V. O. (1992). *Pursuit of Truth*. Cambridge MA: Harvard University Press.
Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* 104: 1436–1441.
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.
Schupbach, J. N. (2017). Inference to the best explanation, cleaned up and made respectable. In T. Poston & K. McCain (eds.) *Best Explanations: New Essays on Inference to the Best Explanation* (pp. 39–61). Oxford: Oxford University Press.
Sterelny, K. & Griffiths, P. (1999). *Sex and Death*. Chicago: University of Chicago Press.
van Brakel, J. (1986). The chemistry of substances and the philosophy of mass terms. *Synthese* 69: 291–324.
van Brakel, J. (2005). On the inventors of XYZ. *Foundations of Chemistry* 7: 57–84.
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.
Weisberg, M. (2005). Water is not H_2O. In D. Baird, E. Scerri, & L. McIntyre (eds.) *Philosophy of Chemistry: Synthesis of a New Discipline* (pp. 337–345). New York: Springer.
Williamson, T. (2018). *Doing Philosophy*. Oxford: Oxford University Press.