Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-022-00613-5

Refining the Bayesian Approach to Unifying Generalisation

Nina Poth[1]

Accepted: 5 January 2022

Abstract

Tenenbaum and Griffiths (*Behavioral and Brain Sciences* 24(4):629–640, 2001) have proposed that their Bayesian model of generalisation unifies Shepard's (*Science* 237(4820): 1317–1323, 1987) and Tversky's (*Psychological Review* 84(4): 327–352, 1977) similarity-based explanations of two distinct patterns of generalisation behaviours by reconciling them under a single coherent task analysis. I argue that this proposal needs refinement: instead of unifying the heterogeneous notion of psychological similarity, the Bayesian approach unifies generalisation by rendering the distinct patterns of behaviours informationally relevant. I suggest that generalisation as a Bayesian inference should be seen as a complement to, instead of a replacement of, similarity-based explanations. Furthermore, I show that the unificatory powers of the Bayesian model of generalisation can contribute to the selection of one of these models of psychological similarity.

Keywords Generalisation · Similarity · Bayesian inference · Unification · Mutual informational relevance

1 Introduction

The problem of generalisation is that of explaining how humans and other animals are capable to generalise their behaviour from old to novel instances. For example, what enables us to treat apples and pears as fruit, but not mushrooms? One answer to this question is, roughly, that the apple and the pear are more similar to each other than the mushroom is to either of them, and so it makes sense to treat the apple and the pear, but not the mushroom, as fruit.

Some psychologists (e.g., Rosch and Mervis 1975; Smith et al. 1984; Sloman and Rips 1998; Hahn and Ramscar 2001) and philosophers (e.g., Shea 2014; Gärdenfors

✉ Nina Poth
nina.poth@rub.de

[1] Department of Philosophy II, Ruhr-Universität Bochum, Bochum, Germany

Published online: 18 February 2022

2000; Decock et al. 2016; O'Brien and Opie 2004) rely on well-defined notions of similarity to answer a broad range of questions about such generalisation problems. However, at least some of these notions are classically opposing. One popular example is the dispute between Shepard's (1962, 1987) geometric-distance model and Tversky's (1977) feature-matching model of similarity. As far as such similarity-based explanations are theoretically disconnected, it is currently unclear to what extent they can render generalisation as a unified psychological phenomenon.

Other psychologists such as Tenenbaum and Griffiths (2001) (henceforth T&G) argue that principles of Bayesian inference provide a more basic explanation of both similarity and generalisation (see also Kemp et al. 2005; Lake et al. 2015; Ullman and Tenenbaum 2020; Austerweil et al. 2019; Tenenbaum et al. 2011; Kemp et al. 2012). The idea is that we are more likely to treat the apple and the pear, as opposed to the mushroom, as fruit because they are more likely to originate from a common 'fruit' concept. The Bayesian approach has generated novel hypotheses and experimental investigations in domains such as word-learning and language development (Xu and Tenenbaum 2007), communication (Frank et al. 2009) and causal learning (Gopnik et al. 2004; Tenenbaum et al. 2006), to name but a few. The rising popularity of Bayesian models in computational psychology and cognitive science illustrates a tendency to substitute the earlier notion of similarity with the notion of Bayesian inference of concepts in such explanations. For instance, T&G propose that the greatest virtue of the Bayesian approach is that it unifies the heterogeneous approaches to similarity and, thereby, best explains generalisation as a psychological capacity underlying similarity judgments, but they do not defend this specific proposal in a rigorous way.

The problem that this debate poses is that both Bayesian inference and similarity remain key concepts in psychological explanations of generalisation, but it is unclear how they fit together. How do Bayesian analyses stand to similarity-based explanations of generalisation? Do Bayesian analyses supersede similarity-based explanations of generalisation? In what sense do these approaches align in their explanatory targets (i.e., inferences structured by similarity representations or probabilistic inferences of concepts)? In how far does the Bayesian approach provide an 'alternative' to similarity-based explanations of generalisation? To answer these questions, conceptual clarifications are needed and a systematic investigation of the basic assumptions associated with these frameworks.

This paper aims to contribute to a better understanding of the relationship between similarity-based and Bayesian explanations of generalisation. In a first step, T&G's Bayesian approach will be systematically analysed and it will be clarified in which sense it attempts to unify Shepard's and Tversky's classically opposing approaches to similarity and generalisation. This analysis reveals that the earlier notion of similarity remains important, alongside Bayesian inference, to explain the psychological capacity of generalisation. In a second step, it will be shown that the implicit reliance on similarity raises issues for the explanatory status of the Bayesian unification, which only reconciles Shepard's and Tversky's specific definitions of similarity by virtue of its agnosticism concerning assumptions about what similarity is. I subsequently propose a novel way to understand the unificatory contribution of the Bayesian model of generalisation to similarity-based accounts. Specifically, I argue that an additional virtue to the mathematical elegance and broad scope of T&G's approach is that it unifies distinct aspects of generalisation in a novel sense that has yet received only a little attention in cognitive science: following Myrvold's (2003, 2017) criterion, T&G's approach renders these distinct aspects mutually informationally relevant under a set of coherent probabilistic assumptions. I argue that this unification is complementary to the earlier similarity-based explanation, and it fulfils a heuristic role in the process of choosing between them for further inquiry into the psychology of generalisation. Specifically, building on Myrvold's approach to unification and an approach to Bayesian model selection by Colombo and Hartmann (2017), I argue that the Bayesian unification justifies a choice of the geometric approach to similarity as a basis for a computational explanation of generalisation as an effect of internal similarity computations. In this sense, T&G's model functions as a case study illustrating a successful unification in cognitive science.

Before proceeding, it is worth clarifying the terms 'generalisation', 'similarity' and 'concept'. By 'generalisation,' I mean a pattern of behaviour that expresses the capacity to treat one object in the same way as another. One can treat objects more or less like another, and in this sense, generalisation is a graded phenomenon. In psychological experiments, the degree of generalisation is often measured by subjects' relative frequency to confuse a set of stimuli, compared to all trials of an experiment, but it can also be assessed by recording explicit similarity judgements, whereby subjects rate the similarity of two items on some scale (e.g., a Likert scale). By 'similarity,' I mean a mental representation of the relation between a set of items. In the current context, this relation is expressed in terms of the items' shared or distinct attributes or their geometric distance in a psychological space. It can be expressed in other ways, for example in terms of the transformation distance (Hahn et al. 2003). Finally, by 'concept,' I mean a mental representation of instances that fall under a category, and this can (but does not have to) be a function of their similarities. Following the common doctrine in cognitive science, concepts are the constituents of thought and crucial for categorisation, inference and learning (Margolis and Laurence 1999). In the current context, concepts constitute the contents of hypotheses in Bayesian inference.[1] Finally, these aspects hang together: generalisation patterns correspond to the outputs of (internal) computations over similarity representations or concepts.

The structure of this paper is as follows. In Section 2, I outline the apparent conflict between Shepard's and Tversky's models of similarity and generalisation. In Section 3, I assess T&G's Bayesian approach to generalisation and concept learning, and its relationship to similarity-based explanations. In Section 4, I introduce Myrvold's informational-relevance criterion, based on which I justify the additional unificatory power of T&G's approach in Section 5. In Section 6, I discuss one possible contribution of this result to the selection of a similarity-based model. Section 7 summarises my findings and provides a brief conclusion.

[1] How exactly concepts are structured has been the focus of much debate in cognitive science (Margolis and Laurence 1999). The Bayesian approach endorsed in this paper is not committed to any specific view on this matter. Readers who adopt a specific view on concepts such as the prototype view, the theory-theory view or conceptual atomism are free to adopt their preferred choice.

2 Two 'Opposing' Approaches to Generalisation and Similarity

Although T&G describe Shepard's (1987) *Universal Law of Generalisation* and Tversky's (1977) *feature-matching model of similarity* as "classically opposing" (Tenenbaum and Griffiths 2001, p. 336), they are short on explaining in which way. To motivate the need to unify these approaches in the first place, I want to call attention to their different mathematical assumptions and empirical contents.

2.1 Different Mathematical Assumptions

The most fundamental assumption of Shepard's model is that dissimilarity is a function, δ, that maps pairs of points, a, b, in a geometric space to a non-negative number, in a way that satisfies the metric axioms.

1. **Minimality**: $\delta(a, b) \geq \delta(a, a) = 0$, where $a \neq b$.
   The distance between an object and itself is zero, and is smaller or equal to the distance between two non-identical objects.
2. **Symmetry**: $\delta(a, b) = \delta(b, a)$.
   The distance between two objects must be the same, regardless of the direction in which it is measured.
3. **Triangle inequality**: $\delta(a, b) + \delta(b, c) \geq \delta(a, c)$.
   In a triadic comparison, one distance must always be shorter or equal to the sum of the other two distances.

Shepard (1962, 1987) models the perceived similarity of two objects as the inverse of their geometric distance. This means that the objects are more similar to each other whenever their vectors are located closer in similarity space (relative to a fixed set of dimensions). Correspondingly, the first axiom says that the similarity between an object (e.g., a specific apple) and itself must be greater or equal to the similarity between two distinct objects (e.g., the apple and a banana). Following the second axiom, the similarity between the apple and the banana must be the same as the similarity between the banana and the apple; following the third axiom, the similarity between the apple and a pear must be less than the joint similarities of the two pairs {apple, banana} and {banana, pear} (unless the banana lies on a straight line between the apple and the pear in which case the similarity between the apple and the pear is equal to the joint similarities between {apple, banana} and {banana, pear}).[2]

In contrast, Tversky's approach relies on set-theoretic relations; these are independent of the metric axioms. A key assumption of this approach is that objects are decomposable into sets of discrete features that take the form of atomic symbols. The *ratio model* represents the similarity between two objects, a and b, as the ratio of the set of their common features to the sum of their common features, the weighted set

[2] There are multiple ways to measure geometric distance. Two prominent examples are the city-block metric and the Euclidean metric. See Gärdenfors (2000, pp. 18-24) for details and discussion.

of features distinct to a and the weighted set of features distinct to b. Formally:

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}, \text{ for some } \alpha, \beta \geq 0, \qquad (1)$$

where $f(A \cap B)$ represents the features common to a and b, $\alpha f(A - B)$ represents the weighted features distinct to a and $\beta f(B - A)$ represents the weighted features distinct to b. A judgement of similarity, S, is a linear function of sets of common and distinct features, and the output of this function depends on the direction of the comparison. S increases linearly with an increase in the number of common features and decreases linearly with an increase of the number of distinct features. When $f(A - B) \neq f(B - A)$ and $\alpha \neq \beta$, then changing the order of the distinct sets of features with fixed weights in the model accommodates directionality, such that $S(a,b) \neq S(b,a)$. Similarity becomes non-directional when either of these conditions is not met, so that $S(a,b) = S(b,a)$ whenever $\alpha = \beta$ or $f(A - B) = f(B - A)$.[3]

2.2 Different Empirical Contents

Aside from their mathematical differences, the two models make conflicting predictions about how subjects (implicitly or explicitly) judge the similarity of pairs of stimuli, and conflicting predictions about how probable subjects are to generalise their behaviour towards them follow. Especially interesting are two distinct phenomena, which I will refer to as ‘the exponential gradient’ and ‘directionality’.

The exponential gradient has become popular as the ‘Universal Law of Generalisation’ (Shepard 1987, henceforth ‘ULG’). It states that the probability of an agent to generalise her behaviour (e.g., eating) from one stimulus (e.g., an apple) to another stimulus (e.g., a pear, or a banana) is an exponentially decreasing function of the perceived stimulus dissimilarity. Specifically, the psychophysical function characterising this effect maps confusion probabilities[4] onto distances in a geometric space. While geometric distances model subjects’ internal representations of stimulus dissimilarity, it is assumed that dimensions model perceptual qualities, such as the loudness of tones or the sweetness of fruit (Shepard 1962; Gärdenfors 2000). The mapping is such that pair-wise distances in the geometric model are negatively exponentially related to pairwise confusion probabilities. It is a robust result that this mapping predicts generalisation patterns across a wide range of different conditions, including different stimuli (e.g., colours, tones, geometric shapes, Morse code signals), modalities (e.g., loudness or brightness), and species (e.g., mammals and birds; insects, fish, amphibians, reptiles (Ghirlanda and Enquist 2003, Cheng 2000)), Thus, due to its generality, it is commonly understood as a psychological law.

[3] Other versions of feature-matching exist; most famously, Tversky’s contrast model. See Tversky (1977) for details.

[4] A confusion probability can be measured as the percentage of errors associated with ‘same’ or ‘different’ judgements for a stimulus pair in an experiment. For example, if out of 100 trials a pair of two different stimuli, $\{a, b\}$, is judged ‘same’ in 40 trials, then the confusion probability associated with that pair is 40.

However, there is an important aspect of generalisation that the law seemingly fails to predict and that may violate some of its basic assumptions. In particular, Tversky's (1977) charge against Shepard's approach to similarity is that it cannot predict effects of directionality, which occur when the pair (a, b) is judged more or less similar than the pair (b, a). A popular example is Tversky's (1977) observation that people typically judge Tel Aviv as being more similar to New York than vice versa. Such an effect is surprising under the geometric model: if asymmetries correspond to directionality effects, then if the symmetry axiom would accurately describe how people judge similarities, we should not expect people to display directionality in how they judge similarities. Experimental evidence for directionality in human similarity judgements has been found for both perceptual and conceptual stimuli. For instance, Tversky and Gati (1978) conducted a set of studies on undergraduate students whereby they tested similarity judgements on sets of schematic faces, cities and countries. Additionally, Tversky (1977) found significant effects of directionality in Rothkopf's (1957) Morse-code data and Krantz and Tversky (1975) report evidence for directionality in similarity-judgements of rectangles. Furthermore, Krumhansl (1978) identifies evidence for directionality in the results of a study by Rosch (1975), who tested similarity judgements of focal colours, and in a study by Rips (1975), subjects were more likely to attribute a disease to an atypical species if the typical species carried the disease than to attribute the disease to a typical species if the infected species was atypical. More recently, Hahn et al. (2009) provide further evidence that similarity judgements are directional for animations morphing one object into another from the same basic category. These findings suggest that directionality effects are robust within the domain of human similarity judgement and inductive generalisation.

Tversky explains these effects of directionality based on his alternative feature-matching approach, according to which the same stimuli are sometimes judged to be differently similar depending on their order of presentation and the salience associated with their shared and distinct features. In the above example, New York is commonly judged to be less similar to Tel Aviv than vice versa, for the following reasons. First of all, these cities differ in how prominent they are, such that New York can be associated with more distinct features than Tel Aviv. This introduces an initial difference between the cardinalities of their sets of distinct features, so that $f(A - B) < f(B - A)$ in Eq. 1. Secondly, one can introduce a difference in the weights associated with these sets of features, so that $\alpha > \beta$, which Tversky justifies with the assumption that the first and second stimuli of a comparison take different grammatical roles in directional tasks of the form 'how similar is a to b?' (as opposed to non-directional formulations like 'how similar are a and b to each other?'); the first stimulus plays the subject, and the second plays the object. Tversky assumes that subjects pay more attention to the subject than to the object; hence, $\alpha > \beta$.

Under these conditions, directionality can be derived with the ratio model. When holding fixed the cardinalities of the sets of common and unequal distinct features and associated unequal weights in Eq. 1, reversing the order of the comparison will evoke a change in similarity. Correspondingly, Tel Aviv is judged to be more similar to New York than vice versa because, in the first direction, New York's distinct features are weighted less, so that the ratio of the common to distinct features is overall greater than in the second direction. The directionality effect can be 'cancelled' by eliminating the initial inequalities, either those associated with the cities' prominence or those associated with their grammatical roles. Due to this flexibility, the model accommodates a variety of other data sets that illustrate directionality effects, including Morse-code signals, countries, faces and geometric objects (Tversky and Gati 1978).

It should be noted that the geometric model can be expanded to in a sense accommodate directionality effects. For instance, Nosofsky's (1986) biased-choice model represents the dissimilarity between two objects as a sum of weighted geometric distances, $d_{ij} = \left[\sum w_k |x_{ik} - x_{jk}|^r\right]^{1/r}$, where $r = 1$ for the city-block distance and $r = 2$ for the Euclidean distance, and w_k is the weight, w, attached to a dimension, k, and $0 \leq w_k; \sum w_k = 1$. Increasing and decreasing w_k expands and shrinks the metric scale, and so an initial distance between two points can be lengthened or shortened. Nosofsky (1986, pp. 54-55) thereby dissociates "similarity representations", which remain subject to metric constraints, from additional "rather complex attention and decision processes" that operate on them by adding a weighting factor. However, a worry with this approach is that overall, similarity might no longer monotonically relate to metric distance because directionality is only accounted for by the addition of other processes, which makes it a trivial generalisation in the sense that any data can be accommodated by any model by just adding assumptions to the model. However, this does not show that the desired effect (i.e., directionality) comes out of the core assumptions of the approach, which in the case of the geometric approach it does not. Furthermore, accommodating changes in similarity judgements of objects, depending on their order in this way in a geometric model does not seem to capture the relevant sense of directionality because this derivation does not show that similarity becomes asymmetric. Within the same perceptual space, the distance remains the same. If what explains directionality effects is only the addition of weights to the model, but not the fact that similarity is represented as a geometric distance function, then this seems to not add justification for why similarity should be modelled as a geometric distance function, as opposed to a function of weighted sets of features, which is the actual point of dispute.[5]

2.3 Can generalisation be analysed as a unified psychological phenomenon?

Taken together, the two frameworks are classically opposed to each other because they think about similarity in mutually incompatible ways, and they generate different empirical predictions about how people generalise their behaviour depending on the similarity of the objects targeted to this behaviour. Recall that Shepard models dissimilarity in terms of geometric distance, and his ULG is a mathematical function that maps the empirical generalisation probability onto geometric distance, predicting that objects that are more similar are exponentially more likely to be confused by subjects (i.e., the exponential gradient). In contrast, Tversky models similarity as a linear

[5] I am grateful to two reviewers for pointing me to this issue.

function of the set-theoretic overlap of weighted features. Under the assumption that subjects sometimes focus more on the second stimulus shown in a comparison than the first, the feature-matching model accommodates the finding that people will sometimes be more likely to generalise their behaviour from one object to another than vice versa (i.e., directionality).

This shows that the two approaches are isolated from each other in the way they predict different aspects of generalisation. On the one hand, the geometric account of similarity implies that similarity is non-directional because it adheres to the symmetry axiom. Therefore, it does not (without additional assumptions) accommodate directionality. On the other hand, the set-theoretic approach assumes that similarity is a linear function of shared and distinct features. Therefore, it does not accommodate the observation that the objective generalisation probability decreases exponentially with a decrease in similarity.

This poses a problem: Given that two distinct approaches are required to derive both predictions, it is not clear whether the two effects (i.e., the exponential gradient and directionality) express a common underlying cognitive capacity. This is because the disunity between the distinct definitions of similarity disconnects the evidence for directionality from the evidence for the exponential gradient. Given this disconnection, there is currently little reason to assume that these distinct phenomena result from the same underlying process of computing similarities, since they are associated with two theoretically disconnected similarity-bases. Given either account, observing an exponential generalisation gradient gives us no reason to expect effects of directionality and vice versa. However, the hypothesis that the exponential gradient and directionality do not express one and the same cognitive capacity challenges the claim that the psychology of generalisation studies a unified cognitive phenomenon.

So where to go next? Both models derive a large variety of phenomena about generalisation, and a choice between these two different accounts is unavailable on objective grounds, for none of them is clearly better supported by the total evidence than the other. Should one give up on a unified study of generalisation?

3 T&G's Bayesian Approach to Generalisation

T&G (2001) protest that generalisation can be analysed as a unified psychological phenomenon. Specifically, they claim that their Bayesian approach to generalisation can reconcile the geometric and feature-matching approaches under a unified explanation of similarity and generalisation that is independently empirically-supported, suggesting that their Bayesian approach should be preferred to explain these phenomena. They put this claim forward along the following lines.

> Here we recast Shepard's theory in a *more general* Bayesian framework and show how this naturally *extends* his approach to the more realistic situation of generalizing from multiple consequential stimuli with arbitrary representational structure. Our framework also *subsumes* a version of Tversky's set-theoretic model of similarity, which is conventionally thought of as the primary

> alternative to Shepard's continuous metric space model of similarity and generalization. *This unification* allows us not only to *draw deep parallels* between the set-theoretic and spatial approaches, but also to significantly *advance the explanatory power* of set-theoretic models. Tenenbaum and Griffiths (2001, p. 629, emphasis added)

The phrases that are highlighted in this quote suggest various ways in which T&G's proposal can be understood; so it remains unclear what this unification precisely amounts to and why it makes T&G's Bayesian approach explanatorily powerful. Philosophers have valued unifying explanations because they reduce the number of independent phenomena that have to be accepted, thereby making the explanatory target simpler and more comprehensible (Friedman 1974). Furthermore, unifying explanations have been thought desirable because they obtain a great deal of evidential support (Glymour 1980). Unification is often understood as the ability to derive a large number of distinct phenomena from only a few argument patterns (Kitcher 1989). However, others such as Morrison (2000) and Potochnik (2011) argue convincingly that the term 'unification' has many disparate interpretations that are often separate from the goals of scientific explanation. Generally, there is wide agreement that unifying theories are worth having, but it is an unresolved issue what they amount to.

In light of these debates, the unificatory status of T&G's Bayesian approach calls for qualification and more careful treatment. To this end, the following sections justify this proposal and critically assess its improvement over similarity-based explanations. Three questions are of special interest. Firstly, what is being unified? I will argue that T&G's model unifies the exponential gradient and directionality effects, and not (as they seem to suggest) the mathematical assumptions of the geometric and feature-matching models of similarity. Secondly, which criterion of unification should be used to assess the success of this proposal? I will argue that this unification is achieved by rendering these phenomena probabilistically dependent, by showing that they obey principles of Bayesian inference. Thirdly, should this perspective replace existing similarity-based explanations of generalisation? I will argue that it should not replace similarity-based approaches because its unifying contributions can be seen as being complementary to them.

3.1 Generalisation as a Bayesian Inference Task

At the heart of T&G's model is a single scheme of Bayesian inference that specifies the probability of a hypothesis in light of a piece of evidence following a version of Bayes' theorem:

$$Pr(H|E) = \frac{Pr(E|H)Pr(H)}{\sum_{H' \in \mathcal{H}} Pr(E|H')Pr(H')},$$

where the posterior is the ratio of the product of the likelihood and prior associated with H to the sum of the products of likelihoods and priors for all alternative hypotheses, H'.

This scheme can be used to analyse generalisation tasks of the following exemplary form. *An agent sees a mushroom with a red head and white spots, and infers that this mushroom is probably an instance of the concept* FLY AGARIC MUSHROOM. *Upon observing another mushroom with a red head and white spots, they must decide: given that the first mushroom is probably an instance of the concept* FLY AGARIC MUSHROOM, *how probable is it that the second mushroom is an instance of this concept as well?* It is apparent from this example that T&G think of the generalisation task as one related to concept learning, where the inference about what concept a set of instances belong to delivers the grounds to treat these instances as the same kind of thing.

Following Bayes' Theorem, learners should approach this task by combining the likelihoods and priors to yield a posterior associated with a hypothesis about what is the right concept that these pieces of evidence fall under. This task requires a hypotheses space, $\mathcal{H}$, with a set of (possibly infinite) hypotheses, H_1, H_2, ... Each hypothesis is a statement about what concept, C, a set of objects, x, y, ..., fall under; its associated probability value expresses how strongly it is believed. As such, a hypothesis expresses a belief (e.g., that there are edible mushrooms). To formally capture the hypothesis that x is an instance of C, we write $H : x \in C$. The task of concept learning is to find the hypothesis associated with the highest posterior probability or a subset of hypotheses weighted by their posterior probabilities.

The prior probability function, $Pr(H : x \in C)$, assigns a probability value to each hypothesis without regard to the given instances. Which particular form the priors should take in T&G's model is an unresolved challenge (see Tenenbaum and Griffiths (2001) and Shepard (1987) for details on the derivations of the exponential gradient with different priors). However, here I follow the common doctrine in Bayesian cognitive science and conceive of learning as the updating of the old probability value associated with a hypothesis (e.g., about what is the correct concept) to a new value that is obtained in light of new evidence (e.g., new instances for the concept). Then what explains the most interesting aspects of concept learning is the likelihood term, which formally relates the evidence to the hypothesis. For the likelihood associated with the observation of x, given that the hypothesis that x is an instance of C is true, we write $Pr(x|H : x \in C)$. On T&G's approach, the likelihood term is specified by the size principle.

3.2 The Size Principle

The *size principle* says that the probability of the evidence, given the hypothesis about a concept, C, is proportional to the ratio of the size of the concept. This means that the evidence becomes more probable in light of hypotheses about concepts that encompass relatively small sets of entities. The size principle also states that this tendency increases exponentially with an increase in the number of instances, n, that have been observed for the concept. Formally:

$$Pr(x|H : x \in C) \propto \left[\frac{1}{|C|}\right]^n, \tag{2}$$

In Eq. 2, the likelihood of observing x given that x is sampled truly from C is proportional to the inverse of the size of C, raised to the power of n examples.

How concept size should be defined is controversial. T&G assume that it is a function of the concept's extension (i.e., the number of entities falling under the concept). Other authors (Gärdenfors 2000; Poth 2019) suggest that it is a function of the concept's intension (e.g., the possible respects an instance could be like). The following is compatible with both interpretations. I use terms like 'small concept' and 'big concept' as shorthands to refer to the size of the set of entities a given concept encompasses. In this sense, the size principle says that how well the hypothesis predicts the evidence is a matter of the size of the set of instances or the region that is identified with C, and hypotheses about small concepts make the evidence more likely. The rationale for this is that inferences about concepts are justified to the extent that the size of the concept matches the variation in the instances observed for it, so if only a few very similar instances have been observed, then this is best predicted by the narrowest concept that they fall under.

A major role of the size principle in T&G's approach to concept learning is to help learners choose among competing hypotheses that are a priori equiprobable. Suppose x is a fly agaric mushroom. If H says that x is an instance of FLY AGARIC MUSHROOM, and H^* says that x is an instance of MUSHROOM, then, following the size principle, H makes x more likely, provided that FLY AGARIC MUSHROOM is smaller than MUSHROOM. Why is H more probable, although both concepts are compatible with x? Because learners should not infer just any concept compatible with the evidence, but one that makes the evidence very likely. The rational relation between the concept's size and the likelihood associated with the hypothesis comes out of an additional assumption in T&G's approach, which is *strong sampling*. Strong sampling says that instances are chosen independently of each other but not independently of the concept that they are chosen from.

A standard random-sampling experiment from Perfors et al., (2011, pp. 306-307) illustrates the idea. Bag A is small and contains a red and green marble; bag B is bigger and contains a red, green and yellow marble. Given these proportions, it is more likely to blindly pick a red marble from A than from B because the probability of picking a red marble given that one reaches from A is .5, while the probability of picking a red marble given that one reaches from B is .33. Now an experimenter randomly samples a sequence of a red, green, red and green marble (with replacement after any pick) from A or B; the subject's task is to guess which bag is the correct one. Considering the sizes of the bags (which are known to the subject—it is just not known which bag the sequence has been taken from), A is correct. It would be surprising to observe this sequence, were the observations in fact independently sampled from B. In that case, one would expect also a yellow marble to occur in the sequence. The lesson to be learned is that smaller concepts allow fewer variation in the outcome of a random sampling experiment; so if we observe a sample with a small variation, it is likely to be a consequence of a small concept.

3.3 The Exponential Gradient Revisited

Equation 2 accounts for the exponential gradient because it predicts that a preference for H over H^* will increase exponentially with an increase in the number of fly agaric mushrooms observed. However, the size principle by itself does not account for generalisation from x to y, it only describes the task of learning a concept from a few examples. What is missing is an account that explains why and under which conditions a learner would generalise a concept that has been inferred for a known instance to a novel instance. To formally capture this problem, T&G (2001) introduce a *generalisation function*. My version of this function is slightly different from theirs to emphasise that hypotheses are statements about the membership of instances to concepts, as opposed to non-conceptual predictions. Formally, the generalisation task can be captured as follows:

$$Pr(H_{y \in C} | H_{x \in C}, E_{x,y}) = \sum_{H:x,y \in C} Pr(H|E) \tag{3}$$

Equation 3 characterises the task of inferring whether y falls under C, given that x falls under C. More specifically, the task is to infer how probable it is that the concept C that y falls under is the same concept C that x falls under. The variable $E_{x,y}$ stands for the total evidence, that is, the observations of x and y jointly taken into account. This evidential statement says that “there is an x and there is a y”. This task is modelled as a function of the sum of the posterior probabilities associated with all hypotheses about C under which these pieces of evidence fall. T&G suitably call the computation of this function “hypothesis averaging”, “because it can be thought of as averaging the predictions that each hypothesis makes about y’s membership in C, weighted by the posterior probability of that hypothesis” (Tenenbaum and Griffiths 2001, p. 631). All predictions about concept membership, $H \in \mathcal{H}$, constitute a mutually exclusive and exhaustive set of possibilities and their posteriors are normalised to sum to 1, and generalisation lies in the interval [0,1][6].

Following Eq. 3, generalisation from x to y becomes more probable with an increase in the number of hypotheses that subsume x and y under the same concepts. Generalisation is perfect (i.e., equal to 1) only if every concept that has x as a member also has y as a member. This result coincides with Shepard’s (1987) conclusion that generalisation probability increases with an increase in the similarity between y and x. Assuming that x and y are each likely to fall under small concepts (following Eq. 2), then holding fixed the sizes of C, the number of concepts that subsume them both is likely to decrease with an increase in their ‘distance’ in psychological space. In other words, the further y ‘moves away’ from x, the less probable it is that they fall under identical concepts. Furthermore, under the assumption of the size principle (2), the generalisation function will have an exponential shape. This means that generalisation decreases exponentially with a decrease in the number of concepts that subsume them both. For example, if x falls under FLY AGARIC MUSH-

[6] Readers are referred to T&G (2001, pp. 630-631) for additional details on the assumptions about the abstract structure of the hypothesis space.

ROOM and y falls under BICYCLE, then there are very few concepts (e.g., THING IN THE UNIVERSE) that subsume them both. Consequently, it is highly improbable that the concept C that y falls under is the same concept C that x falls under, and so there is little reason for a learner to generalise their behaviour from x to y.

Despite the resemblance to Shepard's approach and the exponential form of the function under the size principle, this result is independent of the geometric axioms; it is only constrained by the axioms of probability and set theory. This result is due to the summing of probabilities over the space of hypotheses and not merely by the method for determining the specific content of the hypotheses (i.e., concepts). The content of the hypotheses could be determined by distance in geometric space, sets of features, or in some other way. The important aspect of hypotheses and their relation to probabilities is that they are mutually exclusive, in the sense that if one hypothesis obtains probability .7, alternative hypotheses in $\mathcal{H}$ cannot obtain a greater total probability of .3, since the probabilities assigned to all hypotheses sum up to 1 (a constraint that follows from the third axiom of probability). Adopting the size principle, learners prefer hypotheses about the most specific concept that these things belong to, e.g., "they are both mushrooms", as opposed to "they are both things in the universe". Correspondingly, the inference towards the smallest common concept coincides with the inference towards the most similar instances, but this obtains regardless of whether similarity and concepts themselves are defined in geometric terms or in another way. It obtains due to the preference of hypotheses about smaller concepts, and its exponential increase in strength with multiple examples.

3.4 Directionality Revisited

To encompass Tversky's directionality results, T&G (2001) provide another version of Eq. 3:

$$Pr(H_{y\in C}|H_{x\in C}) = 1/\left[1 + \frac{\sum_{H_{x\in C, y\notin C}} Pr(H,x)}{\sum_{H_{x,y\in C}} Pr(H,x)}\right], \tag{4}$$

where $H_{x\in C,y\notin C}$ stands for a subset of hypotheses which say that x, but not y, falls under C, and $H_{x,y\in C}$ stands for a subset of hypotheses, which say that both x and y fall under C. Subsets of hypotheses are weighted according to their joint probabilities with the evidence, $Pr(H,x)$. It is apparent that Eq. 4 formally resembles the 'inverse' of Tversky's ratio model (Eq. 1); so T&G stipulate that the term $H_{x\in C,y\notin C}$ represents sets of features distinct to x but not y, $H_{x,y\in C}$ represents sets of features common to x and y, and their corresponding probabilities represent feature weights.

A two-step argument can be made for the claim that T&G's model accounts for directionality effects. The first step is to see that a change in the direction of the inference can produce a change in the weighted probabilities associated with distinct subsets of hypotheses. Consider what happens when Eq. 4 is reversed:

$$Pr(H_{x\in C}|H_{y\in C}) = 1/\left[1 + \frac{\sum_{H_{y\in C, x\notin C}} Pr(H,y)}{\sum_{H_{x,y\in C}} Pr(H,y)}\right], \tag{5}$$

where $H_{y \in C, x \notin C}$ says that y but not x falls under C. In Eq. 5, generalisation depends only on the probabilities associated with hypotheses distinct to y; in Eq. 4, it only depends on the probabilities associated with hypotheses distinct to x. Directionality depends on whether these sets of probabilities are different. Insofar as the probabilities associated with each hypothesis in the generalisation task (i.e., that x falls under C; that y falls under C) differ, a change in the direction of the inference can produce a change in the weighted probabilities associated with these hypotheses. For example, assume that $H_{x,y \in C}$ says that both x and y fall under CITRUS FRUIT and assume that $H_{x \in C, y \notin C}$ says that x but not y falls under PHYSALIS and $H_{y \in C, x \notin C}$ says that y but not x falls under ORANGE. Equations 3 and 1 produce different results whenever the probability that x falls under PHYSALIS but not under ORANGE is different from the probability that y falls under ORANGE but not under PHYSALIS.

The second step is to make explicit that this sort of directionality comes out of the definition of conditional probability.

Definition 3.1 (Conditional probability.) The probability of one proposition, A, to be true, conditional on the truth of another proposition, B, is a ratio of their joint probabilities and the unconditional probability of B. Formally:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}; \; Pr(B) > 0.$$

Conversely,

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)}; \; Pr(A) > 0.$$

Definition 3.1 reveals that conditional probabilities are directional, depending on differences between the constituting unconditional probabilities. For example, the probability that a dice lands 3, given that it lands odds, is 1/3, but the probability that a dice lands odds, given that it lands 3, is 1. But the probability that a dice lands odds, given that it lands even is zero, and so is the probability that it lands even, given that it lands odds. The order in which these events are related in the conditional is relevant to a difference between the conditional probabilities if and only if their unconditional probabilities are different, so $Pr(A|B) \neq Pr(B|A)$ if and only if $Pr(A) \neq Pr(B)$. This illustrates that the link to directionality in T&G's generalisation function consists fundamentally in its probabilistic structure. Following definition 3.1, the results of Eqs. 4 and 5 will be unequal whenever the unconditional probabilities, $Pr(H_{x \in C, y \notin C})$ and $Pr(H_{y \in C, x \notin C})$, are unequal, and symmetries will occur whenever these probabilities are the same.

Taking stock, the foregoing sections have introduced T&G's Bayesian approach to concept learning and shown how its central ingredients, the size principle (Eq. 2) and the generalisation function (Eq. 3) accommodate both the exponential gradient and directionality. The next section reviews the unificatory status of this approach, and the sense in which it reconciles the initial dispute between Shepard and Tversky about how similarity is to be defined to explain these generalisation effects. I will argue that T&G's Bayesian approach to concept learning unifies the cognitive task that is common to the two kinds of generalisation effects (i.e., the exponential gradient and directionality), but it is agnostic about issues that concern questions about similarity. Thus, what is reconciled are not the two similarity-based explanations under dispute, but, instead, the two generalisation effects, and these are now traced back to a single underlying capacity of concept learning.

3.5 Compatibility with Shepard's and Tversky's Approaches

At first sight, T&G's approach carries typical characteristics that are sometimes taken to be sufficient for unification: in line with T&G's proposal, the account is mathematically elegant and broad in scope. Instead of two distinct approaches to similarity we only need one approach to generalisation as Bayesian inference to predict the exponential gradient and directionality. To this end, T&G (2001) themselves draw analogies between their unification and Newton's unification of gravity and astronomy. In both cases, the two original frameworks—in Newton's case, Galileo's laws of terrestrial mechanics and Kepler's laws of celestial bodies and in T&G's case, Shepard's ULG and Tversky's feature-matching function—are difficult to reconcile instantly but are subsequently shown to be compatible with a single mathematical framework that subsumes each as a special case. They argue that the gain from this unification is an increase in the mathematical elegance and predictive scope of the resulting theory—in Newton's case, the theory of universal gravitation and in T&G's case, the Bayesian approach to universal generalisation. Thereby, essentially the same predictions about phenomena of gravity and astronomy, in the case of Newton, and about the exponential gradient and directionality, in T&G's case, can be obtained by adding more specific assumptions to the general mathematical theory. This warranted the choice of Newton's theory over competing alternatives that were less elegant and of a narrower scope.

However, the analogy is not perfect. Even if the Bayesian approach is more elegant and broad, this does not necessarily warrant its choice over Shepard's or Tversky's approaches to similarity when explaining the observed effects of generalisation. This is because, in contrast to the similarity-based approaches, T&G's Bayesian approach obtains these typical unificatory features only because it is agnostic about questions concerning the contents of hypotheses, the structure of concepts and their relationship to the evidence. Yet, these questions are still relevant when explaining why and how learners compute the probabilistic inferences of concepts to solve the generalisation task. To understand this point, it is helpful to adopt Marr's (1982) levels-framework as a heuristic. The computational level of analysis specifies the task of generalisation, why it is appropriate and the logic of its potential solution. The level of representation and algorithm identifies the format and contents of representations (e.g., zeros and ones, in the case of a computer) and how these are manipulated by the system to solve the problem. The level of implementation identifies how this solution is physically realised. For Marr, all three levels are important to explain cognition; none of them suffices in isolation. Various revisions of the framework have been proposed (Danks 2008; Love 2015; Hardcastle and Hardcastle 2015; McClamrock 1991; Poggio 2012); and there is wide agreement that there can be multiple levels between the initial ones.

I suggest that the proper domain of T&G's unification lies at the computational level, where generalisation is described as a task of inferring from a known object whether an unknown object falls under the same concept, and the logic of its solution is the strong sampling assumption. This description is agnostic about issues at lower levels because it excludes any particular assumptions about how concepts are represented in learners' minds and about what algorithm computes their sizes and conditional probability functions. Likewise, Shepard's ULG and Tversky's feature-matching function can be seen as residing at the computational level of analysis. Especially Shepard (1987) characterises generalisation as an inductive inference problem for intelligent agents, and the ULG formalises its ideal, invariant, structure. Tversky's feature-based approach focuses less on the problem of similarity judgement as an inductive inference but likewise starts 'top-down' by identifying a set of mathematical axioms that seem to be obeyed by people's performance in similarity-judgement tasks. As suggested by T&G's model, Shepard's initial derivations of the law, based on the assumption of geometric space, can be generalised based on the notion of a consequential subset. This shows that the form of the generalisation problem is independent of the additional assumptions about whether concepts and features have a geometric or set-theoretic structure, thereby broadening the scope of the ULG to accommodate cases outside the domain of a spatial analysis. Thus, at the computational level, we have a single description of the task of generalisation as a Bayesian inference problem that combines the predictions of both the exponential gradient and directionality effects.

However, while T&G's model unifies Shepard's and Tversky's models of generalisation at the computational level, there are at least two things it does not do that motivate refining its contribution to the Shepard-Tversky debate about how similarity is to be defined when explaining the effects of generalisation. Firstly, Shepard's and Tversky's approaches also make additional distinct geometric and set-theoretic assumptions about the structure of similarity representations and concepts that are not entirely brought into agreement by T&G's characterisation of the Bayesian inference task. In particular, the geometric and set-theoretic models disagree in terms of their more specific versions of how concepts (geometric regions or atomic features) and similarity (geometric distance or set-theoretic overlap) can be understood. These versions reflect distinct theoretical perspectives on conceptual thought: the set-theoretic model understands features as conceptual atoms that have no internal structure (Fodor 1998, p. 121) and their set-theoretic 'overlap' constitutes the degree of similarity among objects in their extension.[7] The geometric model reverses this relationship and defines concepts in terms of the dissimilarity among possible members. Blumson (2018, p. 21) argues that these two conceptions of similarity are logically coherent but also independent of each other.[8] Specifically, they are "neither quivalent nor inconsistent, and neither one entails the other" (ibid.).

Importantly, the Bayesian approach remains uncommitted to any of these perspectives. Bayesian norms of reasoning provide the constraints on how probability assignments should be combined coherently, in the manner of Bayes' theorem, but it does not commit one to any specific view about how concepts are defined, and how one measures the consistency between instances and hypotheses about concepts when determining what probability value a specific hypothesis obtains. The characterisation of generalisation as a Bayesian-inference task assumes *some* notion of a concept to specify the content of a hypothesis in the model and *some* notion of concept size to assess how well a hypothesis about that concept predicts the evidence. However, the Bayesian framework is compatible with many different ways to understand a concept and to measure its size to justify the inferences. Indeed, the elegance and broader scope associated with the Bayesian approach arise from the fact that which *specific* understanding of a concept one adopts is irrelevant to derive the exponential gradient and directionality. For instance, some concepts might correspond to discrete features, and others might correspond to geometrically structured regions, or concepts might correspond to none of these, as long as how the inferences about concepts are combined agree with the norms of Bayesian reasoning. Part of the reason why the Bayesian model is compatible with both Shepard's and Tversky's approaches is that it is agnostic about their additional independent assumptions about similarity representations, even if it combines their separate predictions under a single formal framework. This is how the Bayesian model might be said to 'reconcile' this apparent disagreement. (It 'reconciles', rather than 'resolves' this disagreement, because the disagreement reappears, once Shepard's and Tversky's specific commitments are taken back on board.) In this sense, T&G's unification shows that both effects of directionality and the exponential gradient obey regularities that manifest unified principles of Bayesian inference, but these principles are compatible with a variety

[7] In the current context, a feature corresponds to a simple concept. Simple concepts are those such as BLUE or SQUARE, and these relate to complex concepts such as WHALE or HOUSE in the sense that the meaning of complex concepts is composed of the meaning of simple concepts, while simple concepts cannot be decomposed meaningfully into further conceptual parts. In other words, for the current purposes, I treat features as specific kinds of concepts, while the notion of a concept more generally refers to either simple or complex concepts.

[8] Blumson inspects this relation from a metaphysical perspective (where it traditionally concerns the relation between properties and resemblance), but his conclusions concerning the independence of the two conceptions do not depend on whether one is committed metaphysically to similarity and concepts.

of options for specifying a relevant notion of similarity to represent the concepts inferred in the model.[9]

To clarify the distinct commitments of the Bayesian and similarity-based approaches, it is helpful to relax Marr's three-levels analysis and adopt the possibility of having levels between the levels. At the computational level, we have T&G's Bayesian analysis of the task of generalisation and concept learning, which is to compute the conditional probability of some hypothesis given the evidence. Between the level of computational analysis and the algorithmic level, we find the specification of different kinds of similarity representations to individuate hypotheses in terms of a particular structure of concepts. At this 'level', we can locate the geometric and feature matching models of concepts, where similarity and the size of concepts can be measured either in terms of a measure of the overlap of discrete features or in terms of a measure of geometric distance. We currently have no unification of the additional assumptions about similarity representations at this level because the decision about whether a measure of set-theoretic overlap or whether a measure of geometric distance should be used to specify the size of a concept and its relation to instances in the generalisation task is independent of the analysis of the task as an inference problem at the computational level. When asking questions about how concepts are compared to the available evidence to compute the inference, these additional assumptions are not fully reconciled by the Bayesian analysis, since the Bayesian approach resides at the computational level, but these questions concern levels below, specifically questions concerning the individuation of hypotheses about concepts and their relationship to perceptual instances. Once these concerns are re-introduced to explain

[9] That these conceptions are used to make representational claims about concepts, as postulated in explanations of generalisation, is not an uncommon view in the cognitive science literature. For example, Hahn and Chater (1998) analyse generalisation in terms of 'representation-matching', a process in which the criterion for generalisation is the degree to which a previously stored set of items associated with a label matches in similarity to novel items. A novel item should then be assigned the same label as the prototype or exemplar representation it is most similar to. They consider Tversky's set-theoretic and Shepard's geometric models as special cases of matching representations in this way. At the philosophical end, Gärdenfors (2000) develops the geometric conception into a framework of knowledge representation, with concepts modelled as regions in a psychological space that is defined by continuous perceptual quality dimensions (e.g., colour and shape) and with distances among points in a region modelling the dissimilarity between instances of the concept. As a method for concept learning, Gärdenfors (2000, ch. 4) proposes a function that maps regions onto central points to derive a Voronoi tessellation that carves up the phenomenal space into discrete sets of concepts from experience. These authors keep a distinction between similarity, concepts, and generalisation or concept learning to explain how this behaviour could arise as a function computed over assumed internal mental representations. Similarly, Nosofsky (1992, p. 26) views the contribution of the geometric conception as providing "scaling techniques [that] can be viewed as psychological models for the mental representation of interobject similarity", and "the role of discrete-feature and network approaches as components in cognitive process models [as being] of equal importance." It should be noted that the status of these representations seems to be that of scientific instruments to summarise data, perform experimental tests and predict behaviour, and so there is no commitment that this is how people *actually* compute generalisation with concepts in their minds. This resonates with Shepard's (1963) view of the practical advantages associated with the geometric conception, which provides a "spatial 'map' of the stimuli [... that] seems to reveal, in succinct and immediately assimilable form, the latent psychological structure that was contained only obscurely in the initially much larger array of [similarity-judgment] data."

generalisation, there remain two distinct options for how to specify similarity, in terms of either a function of geometric distance or a function of feature matching.

The way in which T&G (2001) understand the Bayesian model as reconciling the geometric and set-theoretic assumptions concerning similarity representations is by identifying concepts with their extension, such that every concept corresponds to a set of points or values representing stimuli[10]. However, there remains a difference between these assumptions that motivates not always defining hypotheses about concepts in this way. An advantage associated with the geometric representation is that it expresses additional useful information about the (rich) structure of concepts that get lost under the interpretation of sets. In particular, the geometric model adds information about the structural relation (i.e., distance) between individual points to approximate the size of a region. Modelling concepts as mere sets of points does not cover this additional information. (The difference is between an ordinal and a ratio scale.) The additional information is useful for modelling concept learners that are non-omniscient and that cannot know the actual number of instances falling under a concept, that is, learners to which the extension is epistemically inaccessible. For example, it is plausible that one cannot know the instances that have in the past, do, and ever will in the future fall under the concept MUSHROOM. Concepts are forward-looking and have an open-ended character in this sense. Since the exact number of mushrooms is not known to the agent who has only limited experience with the instances that fall under a concept, it makes sense to approximate this quantity by representing the relationship *between* individual perceptual experiences. The notion of similarity as a geometric distance is instrumental in modelling these relationships, and it accounts for the possibility of inference under limited knowledge about the concept's true extension. Thus, for reasons of epistemic plausibility, one might not always want to identify a concept with its extension, and so the geometric model remains a reasonable option aside from using only sets of points.

The second point to motivate refining T&G's approach is that it is currently unclear how their model's greater unifying power supports the explanatory advancement that they envision over similarity-based approaches. There is an ongoing debate about whether unifying Bayesian models of cognition bear obvious relations to better scientific explanations than available alternatives. One line of this debate criticises Bayesian models at the computational level for failing to deliver computational details relevant to understanding the psychological mechanisms that underlie the many kinds of behaviours that they predict so elegantly (Bowers and Davis 2012; Jones and Love 2011; Colombo and Seriès 2012; see Griffiths et al. (2012) for a response).[11] Admittedly, similarity-based approaches might be no more explanatory in this sense, but this does not seem to warrant replacing them with a Bayesian approach. Decock and Douven (2011) defend the scientific usefulness of the notion of similarity to explain inductive inferences, against Goodman's (1972) earlier charges

[10] I am grateful to a reviewer for pointing this out to me.

[11] Morrison (2000) makes a similar point to dissociate Newton's unification from an adequate explanation of celestial and terrestrial processes. She demands that such an explanation should go beyond correct, elegant, predictions and provide information about true mechanical causes.

that this notion would be too slippery for that purpose. They (p. 73) consider the geometric and set-theoretic conceptions as “the two best such accounts” to make this notion precise.

Instead of contrasting the explanatory success of Bayesian and similarity-based approaches, it might be more appropriate to see them as complementary. Bayesian unification does not rely on any specific assumptions about similarity. However, Bayesian unification also does not supersede similarity-based accounts because these can be helpful to interpret the notions of ‘hypothesis’, ‘concept’ and ‘size’ in the model. Insofar as it is assumed that generalisation is computed using concepts, the geometric and set-theoretic models (or some other similarity-based alternative) offer possible ways of specifying their semantic contents. Furthermore, if generalisation depends on the number of hypotheses that subsume x and y under the same concepts (as suggested by Eq. 3), then the notion of similarity can be used to identify under what conditions concepts are the same (e.g., concepts that have x as a member might have y as a member if and only if these concepts are sufficiently similar according to some threshold).[12]

An example illustrates this point. Following the size principle (Eq. 2), given an encounter of a portobello mushroom, the hypothesis that the concept EDIBLE MUSHROOM is correct should be preferred over the hypothesis that the concept MUSHROOM is correct in the inference because, intuitively, the size of the concept EDIBLE MUSHROOM is smaller than the size of the concept MUSHROOM. The point about keeping similarity-based accounts to explain generalisation and concept learning is that the notion of similarity specifies in what sense EDIBLE MUSHROOM has a smaller size than MUSHROOM. Intuitively, the concept MUSHROOM subsumes a variety of less similar objects than the concept EDIBLE MUSHROOM, and in this sense, the latter has a narrower extension. Proponents of exemplar and prototype accounts have used the notion of similarity in similar ways to define concepts and their relations to the statistics associated with their perceptual instances. For instance, on the account by Rosch and Mervis (1975), the function of a prototype representation is to maximise the similarity among instances of a concept. Intuitively, this is easier for narrow concepts, which combine on average fewer dissimilar objects than concepts with broader extensions.

Similarity-based accounts also provide useful ways to elaborate why the size principle accounts for learning that is intuitively ‘rational’. In the marbles illustration (Section 3.2), agents should prefer the smaller concept because it better predicts the

[12] One can understand concept size in terms of the area of a region in geometric space or, for complex concepts, the inverse of the number of features combined. But one could also understand a concept as taking the form of a Gaussian probability distribution that is defined by its mean and variance, and one could compute the similarity between two concepts by measuring their probabilistic overlap (e.g., using Kullback-Leibler divergence). However, it is important to keep this conception of similarity distinct from the methodological project of unifying generalisation as a Bayesian inference task. The former specifies a way to represent similarities but the latter is a research method to analyse the task at the computational level. One can be Bayesian about the task analysis while being non-Bayesian about similarity representations. Therefore, defining similarity and concept size in terms of probabilities is not contrary to my claim that Bayesian unification fulfils a role for understanding the psychology of generalisation that can be seen as complementary to similarity-based accounts.

small variation in the red-green sample. But how can one specify the conditions of predictive success when justifying this inference? Bayesian confirmation theory recommends studying the confirmatory relations between the evidence and the hypothesis (see Fitelson 1999, for a review). However, this link is currently missing in the size principle; the right-hand side of Eq. 2 does not explicitly relate instances and concepts.

A notion of similarity is helpful to explicate this relationship. For instance, the degree of confirmation of a hypothesis by some perceptual observation can be assessed based on the degree to which the content of the concept 'matches' the content of the perceptual experience, where matching is a similarity-based condition for the correctness of the concept postulated in the inference (see Brössel (2017) and Hahn and Chater (1998) for two perspectives on this proposal). On this basis, the preference of the hypothesis that EDIBLE MUSHROOM is correct over the hypothesis that MUSHROOM is correct can be justified by citing that the evidential statement "there is a brown-looking, round, short, ... mushroom" is more closely predicted by the hypothesis "there is an edible mushroom", as opposed to the hypothesis "there is a mushroom" since, intuitively, the former two statements are more similar in their meaning than the evidential statement and the latter hypothesis. These examples are not exhaustive, but they illustrate that the greater simplicity and generality of the Bayesian model do not fully compensate for the added value from similarity-based explanations in answering questions about concept individuation (i.e., about the conditions under which one concept is identical or smaller, or matches observations better than another). Thus, their replacement is currently not warranted on the grounds of explanatory unification.

How, then, can one understand the contribution of T&G's unification for investigating the psychology of generalisation? I suggest that neither the apparent lack of integration nor the inconclusiveness of explanatory improvement is fatal for the success of T&G's unification. Several authors, such as Dennett (1987), Zednik and Jäkel (2016), Colombo and Hartmann (2017), and Simon (1977) have emphasised that computational-level analyses can play a useful heuristic role in constraining modelling decisions and discoveries that are typically associated with questions at lower levels in Marr's framework[13]. As McClamrock (1991, p. 187) notes, "we often need to know the function of complex system being analyzed to know what aspects of structure to look at."

I think that a similar role can be attributed to T&G's unificatory framework. T&G seem to assume (but do not defend) such an influence when arguing that "from the standpoint of reverse-engineering the mind and explaining why human similarity or generalization computations take the form that they do, a satisfying theory of similarity is more likely to depend upon a theory of generalization than vice versa" (Tenenbaum and Griffiths 2001, p. 637). In the subsequent sections of this paper,

[13] While Marr (1982, p. 329) also emphasises the importance of the computational-level analysis for investigations at other levels and agrees that the levels "are logically and causally related", he also says that "the three levels are only rather loosely related" (Marr 1982, p. 25). In contrast, recent approaches seem to place a tighter connection between the levels.

I elucidate a novel way in which T&G's unification can be analysed by focusing, not only on its compatibility with the set-theoretic and geometric approaches, but additionally on its evidential support, and I suggest that the value of this unification is that it can facilitate a choice between these approaches. In particular, instead of focusing on simplicity and generality as T&G do, I suggest focusing on Myrvold's (2003, 2017) mutual-informational-relevance criterion, which better captures the sense in which the Bayesian unification of the exponential gradient and directionality contributes to further investigations with a similarity-based model.

4 Mutual Informational Relevance

On Myrvold's (2003, 2017) approach, unificatory power is measured by how well a theory makes a set of phenomena[14] mutually inform each other. He uses a Bayesian notion of information, according to which the degree of information that one proposition, p, yields about another proposition, q, is a matter of how probabilistically dependent q is on p (or vice versa), that is, "how much we learn about whether or not q is true when we learn that p is true" (Myrvold 2003, ibid.)[15].

More specifically, the unificatory power, UP, of a theory, T, associated with two descriptions of phenomena, p_1 and p_2, is a measure of the informational relevance, I, between p_1 and p_2, in light of T, in contrast to how informationally relevant p_1 and p_2 are to each other without regards to T. Formally: $UP(p_1, p_2; T) = I(p_1, p_2|T) - I(p_1, p_2)$ (cf. Myrvold, 2003, Eq. 8)[16]. In particular, the degree to which p_2 is informationally relevant to p_1 is defined in terms of probabilistic dependencies. When a theory, T, *renders* p_2 informationally relevant to p_1, then given T, learning that p_2 is true makes it either more or less probable that p_1 is true. Conversely, T fails to unify p_1 and p_2 when the truth of p_2 makes it no more or less probable that p_1 is true. To contrast the unificatory power of T with an alternative theory, T', the comparative measure, $UP_c(p_1, p_2; T, T') = I(p_1, p_2|T) - I(p_1, p_2|T')$, can be used (cf. Myrvold, 2017, Eq. 7)[17]. For additional formal details, the reader is referred to Myrvold (2003).

[14] Myrvold's original notation involves multiple uses of the variable p, which sometimes refers to either a 'proposition' or a 'phenomenon', a 'hypothesis', a 'sub-theory' or a 'body of evidence'. Henceforth, I use p to represent a proposition describing a phenomenon because the relevant degree-of-belief function, $Pr(\cdot|\cdot)$, takes propositions (i.e., not phenomena themselves) as its arguments.

[15] An equivalent account of unification has been independently proposed by McGrew (2003), although he calls it 'theoretical consilience'.

[16] In his original notation, Myrvold includes a variable representing a theoretical background, b. It is common to assume that the background is already accepted as a rational agent's total knowledge, so that $Pr(b) = 1$. I adopt this assumption for matters of simplicity, and avoid explicit mention of b when explaining Myrvold's proposal to identify unificatory power with a measure of informational relevance or probabilistic dependence.

[17] Myrvold (2003) extends this towards multiple phenomena, but the basic principle of unification is the same. Brössel explicates this measure based on Keynes's coefficient of dependence. He characterises UP as a "deviation from [conditional] independence" (Brössel 2015, p. 522).

To figure out *how much more* T can render two propositions informationally relevant to each other, as opposed to T', one needs to consider the extent to which p_1's and p_2's being mutually informationally relevant under T exceeds the extent of p_1's and p_2's being mutually informationally relevant under T'. This can be measured by UP_c. In terms of probabilistic dependence, what UP_c says is that the degree to which T is more unifying than T' depends on how much more probabilistically dependent p_1 and p_2 are given T, as opposed to how probabilistically dependent they are in light of its theoretical alternative, T'.

Myrvold's (2003) example is the shift from a bare-bones geocentric hypothesis to a heliocentric hypothesis. The important difference is that there are aspects of apparent planetary motion that can be anticipated by the latter but not by the former. On the bare-bones geocentric hypothesis, there is no explanatory relation between a planet's motion on its epicycle and the motion of the Sun, since each planet's epicycle travels along a separate circle near Earth; so features of one planet's apparent motion give little to no information about the features of other planet's apparent motions (i.e., $I(p_1, p_2|h_p) = 0$). In contrast, on the heliocentric hypothesis, all planets orbit around the sun. Due to the additional consideration of the motion of the observer's vantage point from Earth around the Sun, features of one planet's apparent motion become informative about features of the apparent motions of other planets (i.e., $I(p_1, p_2|h_c) > 0$). Therefore, the heliocentric hypothesis unifies statements about the motions of the Sun and statements about the motions of other planets better than the bare-bones geocentric hypothesis (i.e., $UP_c(p_1, p_2; h_c, h_p) > 0$).

5 Unifying Generalisation

In this section, I rely on Myrvold's approach to identify more precisely in what sense T&G's (2001) Bayesian model of concept learning has unificatory merit. Let p_1 be Shepard's observational statement, that the probability of a subject to confuse x and y decreases exponentially with an increase in the dissimilarity between them. Let p_2 be Tversky's observational statement that subjects are sometimes more or less likely to confuse the pair (a, b) than the pair (b, a). Let T be T&G's Bayesian model of concept learning, T' Shepard's geometric model and T'' be Tversky's feature-matching model. From the previous section, we know that if T renders p_1 more probabilistically dependent on p_2 than either T' or T'', then it unifies them better.

First of all, when considered in isolation, p_1 and p_2 seem to be probabilistically independent of each other. Knowing that patterns of generalisation take an exponential shape does not make it more or less probable that patterns of generalisation will be directional. Likewise, knowing that patterns of generalisation behaviours are sometimes directional does not seem to make it more or less probable that generalisation is exponential. Therefore, p_1 and p_2 are mutually informationally irrelevant.

With regards to Shepard's approach, the two descriptions of the exponential and directionality patterns of behaviour remain disunified because the additional assumption that dimensions obtain weights remains informationally irrelevant to the exponential form of the distance function. Changing the weight of the dimension will produce a change in geometric distance between any two points on it, but such a change does not bear upon the exponential form of the generalisation function. In other words, knowing that dimensions obtain feature weights provides no reason to expect that generalisation decreases exponentially with an increase in geometric distance. Furthermore, adjustments in weights and directionality do not naturally follow from adjustments in geometric distance; so it cannot be anticipated that generalisation will be directional if it is an exponential function of geometric distance. Correspondingly, p_1 and p_2 remain probabilistically independent in light of T'.

With regards to Tversky's approach, p_1 and p_2 become negatively probabilistically dependent because the relationship between similarity (i.e., generalisation) and matching features in Eq. 1 is linear; this naturally lowers the expectation that generalisation will have an exponential shape. Correspondingly, given Tversky's model, the truth of p_2 makes it less probable that p_1 is true.

Considering T, the dependencies are different. Recall that T predicts p_1 by replacing the likelihood function in Bayes' theorem with the size principle (2). If we adopt the size principle as a specification of the likelihood function when computing posterior probabilities, we can expect that the sum in Eq. 3 decreases exponentially with an increase in the size of a consequential subset. Furthermore, T predicts p_2, since the posterior probabilities in Eq. 3 are defined as conditional probabilities, and so they are likely to be different from the posterior probabilities in Eq. 5. T renders p_1 positively probabilistically relevant to p_2, since the occurrence of p_1 makes it more likely for p_2 to occur, insofar as p_1 is a feature of the size principle and associated with the likelihood function in Bayes' theorem, which implies the definition of conditional probability and therefore makes directionality likely. The exponential gradient becomes more informationally relevant to directionality under the Bayesian analysis than it was before because now p_2 can be anticipated when p_1 is the case.

Considering Myrvold's (2003, 2017) comparative measure of unification, UP_c, we have:

$$UP_c(p_1, p_2; T'', T') = I(p_1, p_2|T'') - I(p_1, p_2|T') < 0;$$
$$UP_c(p_1, p_2; T, T') = I(p_1, p_2|T) - I(p_1, p_2|T') > 0;$$
$$UP_c(p_1, p_2; T, T'') = I(p_1, p_2|T) - I(p_1, p_2|T'') > 0.$$

Thus, T&G's model unifies p_1 and p_2 better than Shepard's and Tversky's models do because T's unificatory power is always greater than the unificatory power of its competitors.

One might wonder in how far the two empirical propositions are more related in the Bayesian framework than by the addition of a weighting scheme on a geometric space[18] since the exponential gradient follows only from the specific version of the likelihood function, the size principle, which is in a sense symmetric[19] while directionality only follows from hypotheses averaging in Eq. 3, which is subject to conditional probabilities but not to the size principle as such. In response, recall that directionality is built into the Bayesian framework based on the definition of condi-

[18] I am grateful to an anonymous reviewer for asking me to clarify this point.

[19] Although likelihoods can be different for different concepts, it follows from the formal structure of Eq. 2 that, when holding fixed the size of C, having x (or y) in one direction will be equally likely as in the other direction of the comparison.

tional probability, and so directionality comes out necessarily as a result of adopting the Bayesian framework. It is an analytic result. In contrast, adding weights to the dimensions in a geometric model is in a sense ad hoc, since it does not follow from the geometric approach; it just makes the accommodation of directionality work. In other words, adding weights is not a core part of the geometric model, but the definition of conditional probability is a core part of the Bayesian framework. Furthermore, the rationale that motivates the size principle comes out of an analogy to Bayesian confirmation theory, where the confirmatory effect on a hypothesis is commonly greater upon the discovery of surprising evidence (as opposed to unsurprising evidence). Whether E confirms H depends on whether E is more probable conditional on H than unconditionally so that E is more expected if it were known that H was true. E confirms H if E is unexpected without H, but would be expected if H was true. This condition identifies with the size principle. It is surprising to randomly observe a particular fly agaric mushroom since there are many more different kinds of mushrooms to be found. However, finding a fly agaric mushroom would not be surprising if it was known that it was sampled as an example from the concept FLY AGARIC MUSHROOM. Or, at least, it would be a less surprising observation than if it was known that it was sampled as an example from the concept MUSHROOM. This is because the latter concept allows for many more possibilities of a mushroom to be like than this particular fly agaric mushroom. As a consequence of the confirmatory effect of surprising evidence, the observation of the fly agaric mushroom will confirm the hypothesis that it was sampled from FLY AGARIC MUSHROOM significantly more than the hypothesis that it was sampled from MUSHROOM. This coincides with the size principle, according to which a piece of evidence imposes a greater confirmatory effect on a hypothesis about a smaller concept, as opposed to a hypothesis about a larger concept, and generally, that the increase in the degree of confirmation is inversely proportional to the increase in the size of the concept.

However, this only shows that the proportionality component of the size principle (i.e., the fact that smaller concepts are relatively better confirmed by the limited evidence than larger concepts) draws analogies to Bayesian confirmation theory. The exponential component (i.e., the condition that the mapping from concept size to degree of confirmation increases exponentially with an increase in the number of instances) is an empirically motivated assumption, and not part of Bayesian confirmation theory itself. The sense in which the empirical propositions characterising the exponential gradient and directionality are more related in the framework of Bayesian inference is that, if one assumes the Bayesian framework and the exponential component, then one expects both directionality and exponential gradient effects to occur, while one would not do so when adopting only either Shepard's or Tversky's models.

6 A Possible Contribution to Model Selection

The implication for the heuristic value of this approach can be precisely stated by focusing on the indirect confirmation of two competing models that initially obtain equal evidential support. The idea, borrowed from Colombo & Hartmann (2017), is that by having available a Bayesian model that unifies them in some sense, a choice can be made in virtue of the relations between the competing models and the unifying Bayesian model. Let T' and T'' be Shepard's and Tversky's competing models, and T is T&G's unifying Bayesian model. Following Colombo & Hartmann, T' will be indirectly better confirmed than T'' when the following conditions are jointly met.

1. T' is more coherent with T than T''.
2. T is itself better confirmed by the total evidence than either T' or T''.
3. There is an asymmetry in the a-priori plausibility of T' and T''.
4. There is an asymmetry in how well T' and T'' are directly supported by the independent pieces of evidence.

If (1)-(4) are jointly met, then due to the greater coherence with T, T' obtains additional indirect support over T'', despite the lack of greater direct support by the evidence.

There are reasons to think that these conditions are met. Basically, (2) seems to be met, since, due to its unificatory power, T&G's model is overall better confirmed. On Myrvold's approach, unification is inherently related to evidential support, if, on the one hand, the pieces of evidence are independent, so that $I(p_1, p_2) = 0$ (and this is the case in our example), and assuming, on the other hand, they are additive, so that $I(q, p_1) = I(q, p_1\&p_2) + I(q, p_2)$. Under these conditions, Myrvold (2003, p.412) defines the mutual support by a set of pieces of evidence, e_1 and e_2, as follows:

Definition 6.1 (mutual evidential support) $I(T, e_1\&e_2) = I(T, e_1) + I(T, e_2) + UP(e_1, e_2; T)$.

Thus, unificatory power acts in favour of the theory by adding the quantity UP to the degree of support obtained from e_1 and e_2 individually[20]; so the degree to which T unifies the evidence better than competing approaches coincides with the degree to which it is thereby better confirmed. As a consequence, since T&G's model unifies the exponential gradient and directionality more than Shepard's and Tversky's models do, it is overall better confirmed.

Furthermore, (3) seems to be met because T' and T'' are a priori unequally probable, depending on the paradigm one adopts to model similarity representations. It is commonly assumed that perceptual contents have a continuous structure because perceptual experiences seem to have contents that often cannot be clearly differentiated from one another (Beck 2019; Haugeland 1981). For example, the perceptual experience associated with two blue colour shades allows for more fine-grained distinctions than a distinction between the concepts AQUAMARINE and TURQOISE. On

[20] T's obtaining additional evidential support by conjoining the information from e_1 and e_2 is due to a formal correspondence between UP and the probability ratio measure of confirmation: $Pr(H|E)/Pr(H)$ (Keynes 1921). Accordingly, H is confirmed by E if and only if the prior probability of H conditional on E is greater than the prior unconditional probability of H, such that $Pr(H|E) > Pr(H)$. So E confirms H only if E is positively probabilistically dependent on H.

the one hand, continuous dimensions can express such information about *how much more or less* similar a pair of objects is because they have a natural zero point. In contrast, the decomposition of objects into discrete sets of entities that either share a target feature (e.g., being turquoise) or not fails to account for the continuous structure of perception (Barsalou 2008). On the other hand, discrete representations have proven practical to model conceptual similarities and their sensitivity to contextual changes (Shepard and Arabie 1979; Tversky and Gati 1978; Gati and Tversky 1984). For instance, Tversky (1977) explains the greater similarity between Switzerland and Italy within an expanded context including Uruguay and Brazil by appealing to the greater influence of categorical background knowledge about their continental location. Furthermore, Shepard and Arabie (1979, p. 89) find that "continuous spatial representations [...] may not fully and explicitly reveal the discrete or categorical nature of consonant phonemes [...], of kin and other category-specific terms [...], of social structures or even, possibly, of continuously variable stimuli that are nevertheless psychologically 'analyzable' [...]." As these examples illustrate, the preference for either paradigm to model similarity representations might find its origin in what stimulus domain is targeted. For perceptual stimuli, the geometric paradigm is often preferred, but to model conceptual stimuli, the set-theoretic approach might be preferable.[21]

With regards to (4), there is reason to be optimistic, since there seem to be asymmetries in the direct support that Shepard's and Tversky's models receive. Tversky's model is supported by observations of directionality effects, but not by the finding that generalisation has an exponential shape (which is not accommodated). Conversely, finding an exponential gradient intuitively confirms Shepard's model more than finding directionality because Shepard's model predicts the exponential gradient but its extended version merely accommodates directionality. Overall, none of the independent pieces of direct support is clearly worth less than the other. Although the discovery of directionality seems to be less general than the exponential gradient with regards to different species and modalities, it is robust and strongly verified within the domain of human similarity judgement for both perceptual and conceptual stimuli (Tversky 1977, 1978; Gati and Tversky 1984, 1987; Krumhansl 1978; Rosch and Mervis 1975; Rips 1975). At the same time, evidence for exponential gradients abounds in the domain of perception (Shepard 1987; Ghirlanda and Enquist 2003; Cheng 2000; Frank 2018), but it is, to the best of my knowledge, sparse in the domain of language. Although these asymmetries do not exactly resemble the perfect asymmetries required by Colombo & Hartmann, they provide intuitive reasons to think that a more complex account of indirect model selection covers the case at hand.

With regards to (1), the desired imbalance in the coherence-relations between T&G's model and either of the two competing models depends on how coherence is

[21] It is interesting to note the different perspective from Gärdenfors (2000), who sees as a major advantage of the geometric approach that it links the continuous structure of perceptual representations with the discrete structure of conceptual thought.

interpreted. Colombo and Hartmann (2017, p. 472) understand coherence in terms of fruitfulness, “where the fruitfulness of a model measures the number and importance of theoretical and practical results it can produce.” Being coherent with a unifying Bayesian model means being positively influenced in this way by seeing the cognitive task as one of Bayesian inference.

It seems to be the case that Tversky’s directionality is less coherent with T&G’s task description than Shepard’s exponential gradient in this sense. On the one hand, Tversky’s model is not driven by analysing generalisation or similarity judgement tasks as a Bayesian inference problem, and analysing the task in this way seems to furnish no theoretical insight and/or practical application of Tversky’s approach to directionality. This is because these approaches to directionality are unrelated. Feature weights in Tversky’s account change due to the objects’ relative order and prominence, but inferences change due to changes in the unconditional probabilities. As previous applications have already shown, Shepard’s approach is very coherent with the Bayesian analysis, which proves fruitful. The advantage of framing the problem of generalisation in terms of Bayesian inference and strong sampling is that it allows a generalisation of Shepard’s law of generalisation outside the domain of perception, such as word learning (Xu and Tenenbaum 2007).

If this analysis is correct, then on the basis of its greater coherence with T&G’s unifying Bayesian model, the geometric model receives additional indirect evidential support over the competing feature-matching model. This provides reasons to choose the geometric model over the set-theoretic model to specify the conditions of hypothesis individuation in the generalisation task and inference over concepts, while such a choice could not be arrived at by considering the empirical evidence alone.

7 Conclusion

Bayesian models of cognition are often said to unify a variety of different aspects of cognition. In this paper, I have focused on an influential Bayesian model by Tenenbaum and Griffiths (2001) and two aspects associated with the capacity of generalisation—the exponential gradient and directionality effects. Previously, Shepard (1987) and Tversky (1977) had explained these aspects concerning two distinct conceptions of similarity. From a Bayesian perspective, however, these aspects are instances of one and the same Bayesian inference task. Tenenbaum and Griffiths suggest that the virtue of this unification lies in its unbounded scope or simplicity. I have argued that it bears an additional heuristic value in terms of Myrvold’s (2003) informational-relevance criterion of unification.

This novel way of understanding the virtues of generalisation as a Bayesian inference task reveals its complementary relation to similarity-based explanations of generalisation. I have only sketched one possible way in which these relationships can be fruitful, based on their confirmatory import and coherence. Other possibilities may be discovered.

Acknowledgements I am grateful for very constructive feedback on earlier drafts of this paper from Mark Sprevak, Alistair Isaac, Peter Brössel, Nicholas Rebol, Adrian Downey, Krys Dolega and three reviewers of this journal.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by a a Postgraduate Research Award from the School of Philosophy, Psychology, and Language Sciences at the University of Edinburgh and a postdoctoral grant from the Institute of Philosophy II at Ruhr-Universität Bochum.

Declarations

Availability of data and material Not applicable.

Code availability Not applicable.

Conflicts of Interest The author declares no conflicts of interest.

References

Austerweil, J. L., S. Sanborn, and T. L. Griffiths. 2019. Learning how to generalize. *Cognitive Science* 43(8): e12777.

Barsalou, L. W. 2008. Grounded cognition. *Annual Review of Psychology* 59(1): 617–645.

Beck, J. 2019. Perception is analog: The argument from Weber's law. *Journal of Philosophy* 116(6): 319–349. https://doi.org/10.5840/jphil2019116621.

Blumson, B. 2018. Two conceptions of similarity. *The Philosophical Quarterly* 68(270): 21–37.

Bowers, J. S., and C. J. Davis. 2012. Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin* 138(3): 389.

Brössel, P. 2015. Keynes's coefficient of dependence revisited. *Erkenntnis* 80(3): 521–553.

Brössel, P. 2017. Rational relations between perception and belief: the case of color. *Review of Philosophy and Psychology* 8(4): 721–741.

Cheng, K. 2000. Shepard's universal law supported by honeybees in spatial generalization. *Psychological Science* 11(5): 403–408.

Colombo, M., and S. Hartmann. 2017. Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science* 68(2): 451–484. https://doi.org/10.1093/bjps/axv036.

Colombo, M., and P. Seriès. 2012. Bayes in the brain—on bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science* 63(3): 697–723.

Danks, D. 2008. Rational analyses, instrumentalism, and implementations The probabilistic mind: Prospects for rational models of cognition, eds. N. Chater, and M. Oaksford. New York: Oxford University Press.

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

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

Dennett, D. C. 1987. The intentional stance, MIT Press, Cambridge.

Fitelson, B. 1999. The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science* 66: S362–S378.

Fodor, J. A. 1998. Concepts: Where cognitive science went wrong, Oxford University Press, Oxford.

Frank, M., N. Goodman, P. Lai, and J. Tenenbaum. 2009. Informative communication in word production and word learning, Vol. 31. In: Proceedings of the annual meeting of the cognitive science society.

Frank, S. 2018. Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of the National Academy of Sciences* 115(39): 9803–9806.

Friedman, M. 1974. Explanation and scientific understanding. *The Journal of Philosophy* 71(1): 5–19.

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

Gati, I., and A. Tversky. 1984. Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology* 16(3): 341–370.

Gati, I., and A. Tversky. 1987. Recall of common and distinctive features of verbal and pictorial stimuli. *Memory & Cognition* 15(2): 97–100.

Ghirlanda, S., and M. Enquist. 2003. A century of generalization. *Animal Behaviour* 66(1): 15–36.

Glymour, C. 1980. Explanations, tests, unity and necessity. *Nous*: 31–50.

Goodman, N. 1972. Seven strictures on similarity. In: Problems and projects, 1st edn. Indianapolis: Bobbs-Merrill.

Gopnik, A., C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks. 2004. A theory of causal learning in children: causal maps and bayes nets. *Psychological review* 111(1): 3.

Griffiths, T. L., N. Chater, D. Norris, and A. Pouget. 2012. How the bayesians got their beliefs (and what those beliefs actually are): comment on bowers and davis (2012). *Psychological Bulletin* 138(3).

Hahn, U., and N. Chater. 1998. Similarity and rules: distinct? exhaustive? empirically distinguishable?. *Cognition* 65(2-3): 197–230.

Hahn, U., N. Chater, and L. B. Richardson. 2003. Similarity as transformation. *Cognition* 87(1): 1–32.

Hahn, U., J. Close, and M. Graf. 2009. Transformation direction influences shape-similarity judgments. *Psychological Science* 20(4): 447–454.

Hahn, U., and M. Ramscar. 2001. *Similarity and categorization.* Oxford: Oxford University Press.

Hardcastle, V. G., and K. Hardcastle. 2015. Marr's levels revisited: understanding how brains break. *Topics in Cognitive Science* 7(2): 259–273.

Haugeland, J. 1981. Analog and analog. *Philosophical Topics* 12(1): 213–225.

Jones, M., and B. C. Love. 2011. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences* 34(04): 169–188.

Kemp, C., A. Bernstein, and J. B. Tenenbaum. 2005. A generative theory of similarity. In: Proceedings of the 27th Annual Conference of the Cognitive Science Society, 1132–1137: Citeseer.

Kemp, C., P. Shafto, and J. B. Tenenbaum. 2012. An integrated account of generalization across objects and features. *Cognitive Psychology* 64(1-2): 35–73.

Keynes, J. M. 1921. *A treatise on probability*. London: Macmillan.

Kitcher, P. 1989. Explanatory unification and the causal structure of the world Scientific Explanation, eds. P. Kitcher, and W. Salmon. Minneapolis: University of Minnesota Press.

Krantz, D. H., and A. Tversky. 1975. Similarity of rectangles: An analysis of subjective dimensions. *Journal of mathematical Psychology* 12(1): 4–34.

Krumhansl, C. L. 1978. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review* 85(5): 445–463.

Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266): 1332–1338.

Love, B. C. 2015. The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science* 7(2): 230–242.

Margolis, E., and S. Laurence. 1999. *Concepts: core readings*. Cambridge: MIT Press.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

McClamrock, R. 1991. Marr's three levels: A re-evaluation. *Minds and Machines* 1(2): 185–196.

McGrew, T. 2003. Confirmation, heuristics, and explanatory reasoning. *The British Journal for the Philosophy of Science* 54(4): 553–567.

Morrison, M. 2000. *Unifying scientific theories: Physical concepts and mathematical structures*. Cambridge: Cambridge University Press.

Myrvold, W. C. 2003. A bayesian account of the virtue of unification. *Philosophy of Science* 70(2): 399–423.

Myrvold, W. C. 2017. On the evidential import of unification. *Philosophy of Science* 84(1): 92–114.

Nosofsky, R. M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology* 115(1): 39–61.

Nosofsky, R. M. 1992. Similarity scaling and cognitive process models. *Annual Review of Psychology* 43(1): 25–53.

O'Brien, G., and J. Opie. 2004. Notes toward a structuralist theory of mental representation. In: Representation in mind, 1–20: Elsevier.

Perfors, A., J. B. Tenenbaum, T. L. Griffiths, and F. Xu. 2011. A tutorial introduction to bayesian models of cognitive development. *Cognition* 120(3): 302–321.

Poggio, T. 2012. The levels of understanding framework, revised. *Perception* 41(9): 1017–1023.

Poth, N.L. 2019. Conceptual spaces, generalisation probabilities and perceptual categorisation. In M. Kaipainen, F. Zenker, A. Hautamäki,and P. Gärdenfors: Conceptual spaces: Elaborations and applications (pp. 7-28). Springer, Cham.

Potochnik, A. 2011. A Neurathian conception of the unity of science. *Erkenntnis* 74(3): 305–319.

Rips, L. J. 1975. Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior* 14(6): 665–681.

Rosch, E. 1975. Cognitive reference points. *Cognitive Psychology* 7(4): 532–547.

Rosch, E., and C. B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7(4): 573–605.

Rothkopf, E. Z. 1957. A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology* 53(2): 94.

Shea, N. 2014. Viexploitable isomorphism and structural representation, (Vol. 114. In: Proceedings of the Aristotelian Society, 123–144: Oxford University Press.

Shepard, R. N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika* 27(2): 125–140.

Shepard, R. N. 1963. Analysis of proximities as a technique for the study of information processing in man. *Human Factors* 5(1): 33–48.

Shepard, R. N. 1987. Toward a universal law of generalization for psychological science. *Science* 237(4820): 1317–1323.

Shepard, R. N., and P. Arabie. 1979. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review* 86(2): 87.

Simon, H. A. 1977. Scientific discovery and the psychology of problem solving, (Vol. 54. In: Models of discovery and other topics in the methods of science, 286–303. Holland: Dordrecht.

Sloman, S. A., and L. J. Rips. 1998. Similarity as an explanatory construct. *Cognition* 65(2-3): 87–101.

Smith, E., D. L. Medin, and L. J. Rips. 1984. A psychological approach to concepts: Comments on rey's "concepts and stereotypes". *Cognition* 17: 265–274.

Tenenbaum, J. B., and T. L. Griffiths. 2001. Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences* 24(4): 629–640.

Tenenbaum, J. B., T. L. Griffiths, and C. Kemp. 2006. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences* 10(7): 309–318.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022): 1279–1285.

Tversky, A. 1977. Features of similarity. *Psychological Review* 84(4): 327–352.

Tversky, A., and I. Gati. 1978. Studies of similarity. *Cognition and Categorization* 1: 79–98.

Ullman, T. D., and J. B. Tenenbaum. 2020. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology* 2: 533–558.

Xu, F., and J. B. Tenenbaum. 2007. Word learning as bayesian inference. *Psychological Review* 114(2): 245.

Zednik, C., and F. Jäkel. 2016. Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese* 193(12): 3951–3985.