Erkenntnis
https://doi.org/10.1007/s10670-023-00696-1

ORIGINAL RESEARCH

Same but Different: Providing a Probabilistic Foundation for the Feature-Matching Approach to Similarity and Categorization

Nina Poth[1,2]

Received: 21 July 2022 / Accepted: 6 May 2023

Abstract
The feature-matching approach pioneered by Amos Tversky remains a groundwork for psychological models of similarity and categorization but is rarely explicitly justified considering recent advances in thinking about cognition. While psychologists often view similarity as an unproblematic foundational concept that explains generalization and conceptual thought, long-standing philosophical problems challenging this assumption suggest that similarity derives from processes of higher-level cognition, including inference and conceptual thought. This paper addresses three specific challenges to Tversky's approach: (i) the feature-selection problem, (ii) the problem of cognitive implausibility, and (iii) the problem of unprincipled tweaking. It subsequently supports key insights from Tversky's account based on recent developments in Bayesian modeling of cognition. A novel computational view of similarity as inference is proposed that addresses each challenge by considering the contrast class as constitutive of similarity and selecting for highly informative features. In so doing, this view illustrates the ongoing promise of the feature-matching approach in explaining perception, generalization and conceptual thought by grounding them in principles of probabilistic inference.

1 Introduction

The feature-matching approach pioneered by Amos Tversky (1977) is a groundwork for models of similarity and categorization and is often appreciated for its ability to tackle directionality effects and the context-sensitivity associated with similarity-judgements (to be elucidated below). The feature-matching approach in its

✉ Nina Poth
nina.laura.poth@hu-berlin.de

[1] Institute for Philosophy II, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany

[2] Institut für Philosophie, Humboldt-Universität zu Berlin, Berlin, Germany

Published online: 08 June 2023

original form remains influential in a variety of domains, such as machine learning (Rahnama & Hüllermeier, 2020), AI (Krawczak & Szkatuła, 2018; Lake et al., 2017), judgement and decision making (Galesic et al., 2018) and computational psychology (Austerweil et al., 2019; Falkowski et al., 2018; Sanborn et al., 2021). Given its wide use in the cognitive sciences, it is desirable to assess the cognitive plausibility of this approach and to identify the relevant principles and mechanisms involved in feature-matching, especially in light of recent advances in thinking about the nature of cognition.

While it is often assumed in the psychological literature that similarity is an unproblematic foundational concept that explains generalization and conceptual thought (Goldstone & Barsalou, 1998), long-standing philosophical problems, most famously raised by Nelson Goodman (1972), challenge this assumption based on triviality arguments that render similarity uninformative in the sense that anything is similar to anything else in some respect (e.g., my foot is similar to the table in front of me in that they are equally distant from the moon). How is one to choose the relevant respect? A promising answer is that similarity derives from processes of higher-level cognition, including inference and conceptual thought (Sloman & Rips, 1998). However, on pain of circularity, it remains unclear how similarity representations can simultaneously be basic to such processes (Decock & Douven, 2011). Existing responses to the conundrum focus on geometric models, pioneered by Shepard (1962), Nosofsky (1986, 1991), & Gärdenfors (2000). According to these accounts, dissimilarities take the form of distances between perceptual representations that are constituted by quality dimensions with a geometric structure. Perceptual representations of individual objects are modeled as points in geometric space, in which concepts take the form of regions, and a cognitive bias can be imposed on similarity in the form of a dimension weighting. These models attempt to solve the conundrum by taking the geometric structure of perceptual space as primitive to similarity representations. An unresolved challenge for these models is to deal with syntactically (e.g., compositionally) structured representations, which are ubiquitous in discussions of object recognition, imagistic cognition, language, knowledge, and more (Langkau & Nimtz, 2010). Although the feature-matching approach is widely regarded as an alternative to the geometric approach, it has received little attention in this debate. As a consequence, the relations between both approaches also remain insufficiently addressed (but see Decock & Douven, 2011 as well as Poth, 2022, for two illustrations).

This paper critically assesses and refines Tversky's feature-matching to offer a new way of approaching these worries. Starting from geometric models of similarity representations in perception, I show how one can incorporate insights from the feature-matching approach if one takes seriously recent developments in Bayesian modeling of cognition. In the critical part of the paper, I argue that the classical version of the feature-matching approach is profoundly flawed since it suffers from three shortcomings: (i) the feature-selection problem, (ii) the problem of cognitive implausibility, and (iii) the problem of unprincipled tweaking. The first problem is that the classical approach does not explain how features are initially selected to judge similarities, since it considers features as trivially given. The second problem is to deliver a plausible notion of feature representation. Tversky initially characterized features as discrete elements, but this conception is highly inadequate for a

similarity-based account of perceptual categorization outside the domain of judgement and voluntary decision-making. The third problem is that feature-matching in its classical form is too flexible; it lacks appropriate rationality constraints to provide a good psychological explanation of why similarity judgements take the form they do. Together, these shortcomings suggest that feature-matching lacks the advantages it initially suggested to have over alternative geometric approaches, which motivates substantial revision of the account.

In the constructive part, I argue that these shortcomings can be addressed with a novel computational account of feature-matching as a Bayesian inference task. The proposed view identifies probabilistic norms governing how feature matching should work, as opposed to providing a causal explanation of how people match features when they judge similarities. Specifically, this account combines three ingredients. The first ingredient is to consider the contrast class as constitutive of similarity. Furthermore, from a probabilistic perspective, an optimal response to a similarity-judgement task is to select for highly informative features. I suggest as a second and third ingredient that rational agents do this by combining a preference for rare features (the second ingredient) with a preference for high-variability features (the third ingredient). This novel understanding of feature-matching addresses the shortcomings of Tversky's approach in the following way: (i) when selecting a feature, agents should consider both, how well its frequency predicts the evidence in a particular context, and how much the range of a feature varies across contexts; (ii) the probabilistic approach is compatible with a notion of similarity that captures the domain of perception and action while understanding features in terms of continuous representations; (iii) similarity effects arise as core phenomena of probabilistic principles, as opposed to unprincipled accommodations of the data.

The implication for the larger debate on the notion of similarity in psychological science and philosophy is two-fold. Firstly, the relationship between probabilistic inference (as a basis for feature-selection) and similarity representation (as a basis for the content of probabilistic inference) is not viciously circular. Secondly, my approach highlights an important insight that was only implicit in both classical feature-matching and geometric approaches to similarity: contextual factors are constitutively relevant to compute similarities.

The structure of this paper is as follows. Section 2 outlines the main tenets of classical feature matching and highlights two major advantages: it accommodates the directionality and context sensitivity associated with similarity judgements. Section 3 outlines three challenges for classical feature-matching. Section 4 responds to these with an alternative account of similarity as Bayesian inference to advance the positive development of feature matching. Section 5 discusses the implications of this proposal for the debate on grounding cognition in similarity representation. Section 6 ends with a brief conclusion.

2 Basic Tenets of the Classical Feature-Matching Approach

On Tversky's (1977) classical approach, similarity is a linear function of set-theoretic overlap. Let $\Delta = \{a, b, c, \dots\}$ be the domain of objects. $A, B, C, \dots$ are the sets of features, where each feature set is associated with an object from Δ (i.e., A

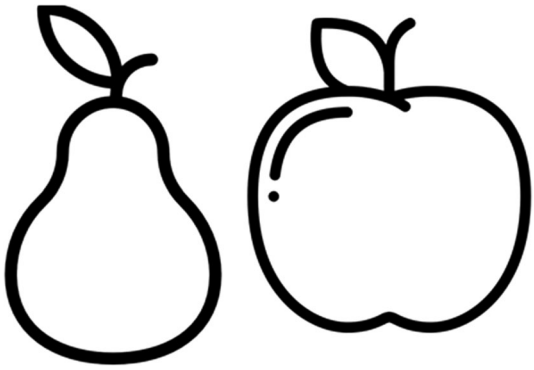

Fig. 1 Representation of a pear and an apple with common and distinct features

represents the features associated with the object a, B represents the features associated with the object b, etc.).

2.1 The Contrast Model

There are many ways to assess set-theoretic overlap with feature-matching. Tversky himself notes that his framework ``encompasses a wide variety of similarity models that differ in the form of the matching function F and in the weights assigned to its arguments'' (Tversky, 1977, p. 333, see also Restle, 1961 and Sjöberg, 1972). He focuses on two variants, the ratio and contrast models. For reasons of space limitations, I focus only on the contrast model, which represents the similarity between a and b, $S(a, b)$, as a linear absolute difference of their shared and distinct features[1] Formally,

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

where $A \cap B$ represents the intersection of shared features, $A - B$ represents the complement of the features that are distinct to a without b, and $B - A$ represents the complement of the features that are distinct to b without a. f is a non-negative scale over a given set-theoretic space and measures the salience of a set of features. The weights, θ, α and β are positive constants between 0 and 1. They determine how much each set of features contributes to the overall measure of similarity. If $\theta = 1$, $\alpha = \beta = 0$, then similarity depends only on the set of common features. Conversely, if $\theta = 0$, $\alpha = \beta = 1$, then similarity depends only on the sets of distinct features.

Figure 1 provides a toy example. Say the two pictures of the objects (take b for the pear, a for the apple) have one feature in common, which is the possession of a leaf, and they differ regarding their shapes, as one of them is round and the other oval. Thus, when holding fixed the weights and scale (i.e., determining each to be 1 in Eq. 1), the two sets of distinct features each obtain a cardinality of 1, and the

[1] The ratio model represents $S(a, b)$ as a ratio of the number of common features to the total number of common and distinct features. See Tversky (1977), for further details.

set of common features obtains a cardinality of 1, and so two distinct features are subtracted from one common feature, resulting in a negative similarity score of -1. This example is highly simplified, as it ignores other features of apples and pears (e.g., sweetness of taste, growing on trees, etc.) and it does not consider differences in their salience. It illustrates the very basic idea of feature matching.

2.2 Accommodating Directionality Effects

A motivation to consider Tversky's approach is that it easily deals with directionality effects. For example, people are more likely to judge Tel Aviv to be similar to New York and they are less likely to judge New York to be similar to Tel Aviv (Tversky, 1977, p. 328). This effect is frequently described in the literature on similarity and categorization studies. Aside from studies confirming the effect by Krantz and Tversky (1975) and Tversky and Gati (1978), Krumhansl (1978) identifies evidence for directionality on the similarities of focal colors. Rips (1975) shows that subjects are more likely to attribute a disease to an atypical species if the typical species carries the disease, as opposed to attributing the disease to a typical species if the infected species is atypical. More recently, Hahn et al. (2009) provided evidence showing that similarity judgements are directional for animations morphing one object into another from the same basic-level category.

The contrast model accommodates directionality under three conditions. Firstly, the task must involve statements of the form 'a is like b' or 'b is like a' (non-directional similarity statements have the form 'a and b are alike'). Assuming a structural correspondence to similarity judgements, Tversky (1977) explains that people judge the similarity between Tel Aviv and New York differently from New York and Tel Aviv because, depending on their order, the terms "Tel Aviv" and "New York" take on different semantic roles in the statements "Tel Aviv is like New York" (S1) and "New York is like Tel Aviv" (S2). Tel Aviv plays the role of the subject in S1 and the role of the object in S2, while New York takes on the role of the object in S1 and the role of the subject in S2. Secondly, one of the distinct sets of features may be more salient, i.e., $f(A-B) \neq f(B-A)$, and salience weights can have an additional influence on directionality. This means that A has a greater or smaller cardinality than B, and therefore a greater or smaller impact on 'how much' dissimilarity is taken away from the overlap in Eq. 1. For example, the distinct features associated with New York might be more salient than those associated with Tel Aviv because New York is, intuitively, more popular and associated with more distinct features. The final condition is that the set of features distinct to the subject in the model are weighted more than the set of features distinct to the object, such that $\alpha \geq \beta$. Regarding the current example, Tversky (1977) stipulates that the distinct features associated with New York are more important when New York plays the role of the object, as opposed to the subject.

When these three conditions hold, similarity judgements become directional in Eq. 1[2] However, it is important to note that changing the relative order of the objects in the comparison (corresponding to a change in the relative position of the sets of distinct features in the model) will not suffice to evoke directionality if the focusing hypothesis does not hold (i.e., when $\alpha = \beta$). To illustrate: when $\alpha > \beta$, then in S1, the set of features distinct to Tel Aviv obtains more weight than the set of features distinct to New York. In S2, the set of features distinct to New York obtains more weight than the set of features distinct to Tel Aviv. Insofar as Tel Aviv is associated with fewer distinct features, its position in second place in the model (corresponding to S2) subtracts a smaller amount of dissimilarity from the common features, which accommodates directionality.

2.3 Accommodating Context-Effects

Another key finding favoring the approach is that similarity is context sensitive. This is illustrated by Goodman's example of an airport checking station, where the respects in which a spectator, pilot and passenger judge pieces of luggage vary, depending on what aspects they attend to—"The spectator may notice shape, size, color, material, and even make of luggage; the pilot is more concerned with weight, and the passenger with destination and ownership. […] Circumstances alter similarities" (Goodman, 1972, p. 445). Examples of context-sensitivity are ubiquitous, but their nature is still not sufficiently explained.

Tversky (1977) explains context effects based on his diagnosticity principle, according to which objects are classified concerning the diagnostic value of their features. Similarity judgements vary with context because a change in context induces a change in diagnostic value, and hence weight, of a feature. Tversky (1977, pp. 28–29) exemplifies this with the feature 'real', which "has no diagnostic value in the set of actual animals since it is shared by all actual animals and hence cannot be used to classify them [… but] acquires considerable diagnostic value if the object set is extended to include legendary animals, such as a centaur, a mermaid, or a phoenix." The replacement of class members alters the diagnostic value of the feature `real' and determines how the remaining objects will be categorized. Equation 1 accommodates this by adjusting feature weights. Features with a higher diagnostic value obtain more weight than features with a lower diagnostic value, thereby changing the output of the similarity function. In the example, 'being real' is weighted higher in the expanded context, which makes legendary animals more dissimilar from other animals as compared to the narrow context.

The significance of these accommodations can be appreciated when comparing feature matching to geometric models of similarity in the tradition of multi-dimensional scaling (Shepard, 1962), which formalize dissimilarity in terms of geometric distance, and therefore inherently rely on the axiom of symmetry. The metric distance between two points, *a* and *b*, is always the same as the metric distance between *b* and *a*, suggesting that similarity judgements are non-directional and independent

[2] It is not required that all three conditions are fulfilled simultaneously. Either the first and the second condition, or the first and the third condition must hold to obtain directionality effects with the contrast model.

of the order in which the objects are compared. Furthermore, geometric models are often associated with universal laws of generalization and perception that highlight aspects of similarity that remain invariant across many different contexts (Shepard, 1987; Sims, 2018). An initial appeal of the classical approach is that its findings raise awareness of these issues.[3]

In what follows, I trade these advantages with diminishing returns. I discuss three challenges showing that directionality effects do not necessarily violate the assumptions made by geometric alternatives, and therefore offer insufficient evidence to refute them. Furthermore, I argue that accommodating context effects is only achieved at a high explanatory cost that can be avoided by recently developed probabilistic models of cognition.

3 Three Challenges for Classical Feature-Matching

3.1 The Feature-Selection Problem

The first problem is that the feature-matching approach fails to offer a plausible story about the choice of features. In any situation, there seem to be innumerably many possible features that could be used to judge the similarity between two objects. On what grounds should agents select the relevant features? For example, 'uncolored' and 'symmetrically shaped' seem irrelevant to judge whether the pear and apple fall under the fruit category. The classical approach focuses on set-theoretic axioms to formalize similarity computations, but does not explain the principles to select features from the objects that have them *before* the matching process. This problem generalizes to non-perceptible objects. To borrow an example from Shanon (1988, p. 309), which features should a person seeing a face choose to describe to another person not seeing it what face she is talking about? There is an innumerable number of features with which she could describe the face. Which are those relevant to both?[4]

Another way to understand this problem is in terms of an implicit circularity. It is often assumed that the problem of selecting properties relevant to an inductive inference task is solved by focusing on the notion of similarity as the primary relationship between objects. Accordingly, the similarity among members of the extension or intension of a property is constitutive for determining its inductive relevance; so similarity defines properties (Decock & Douven, 2011; Goodman, 1972). However, this

[3] Recent advancements in the psychological modeling literature offer some credit to geometric models' context sensitivity, especially in the color domain. Modelers commonly use either the CIELAB or the CIELUV space, but which of these they use depends on the context (e.g., depending on whether the task is to judge the similarity of colors shown on paper or cloth, or shown on a screen). Fitting any particular type of space is very expensive, as it requires a lot of data as input to perform a dimension reduction. Due to these practical difficulties, it is therefore not surprising to not find different spaces for different contexts (e.g., one for olfaction and one for vision). For practical ease, it makes sense to rely on a single kind of space to model a variety of different kinds of data across various contexts. I am grateful to Corina Strößner for pointing me to this possibility.

[4] The problem appears under various guises in the philosophical literature. Bloch-Mullins (2020) and Machery (2007) call it the "feature-specification problem''.

response is not available to proponents of classical feature-matching, who assume that the relevant properties are primary to the similarity relation. The domain of objects, Δ, is presupposed; each object is represented by a set of *given* features. Then, on pain of circularity, similarity cannot be used to define the properties associated with these objects if similarity itself is defined in terms of these properties.

3.2 The Problem of Cognitive Plausibility

The second problem is that the classical feature-matching approach lacks compatibility with an adequate notion of feature representation in perception and conceptual thought. Features are discrete binary entities—an object is either red or not, but it is not more or less red. However, perceptual features have a gradual structure. The perception of color involves no sharp boundaries and varies with lighting conditions, i.e., objects appear more or less red (Beck, 2019). Similar issues arise when identifying features with perceptual concepts, which encode statistical information that indicates how well instances fall under the concept. For example, the perceptual concept RED has focal and peripheral members that differ in their typicality (Douven et al., 2017; Rosch & Mervis, 1975),[5] suggesting that perceptual concepts come in degrees as well. This line of argument expands even to abstract concepts. Consider TEENAGER, which has a range of values that make up this concept's constituting features, e.g., DOOR- SLAMMING and EYE-ROLLING. Bloch-Mullins (2020, p. 617) provokes: "…how forceful should the shutting of the door be, in order to be considered a slam? What is the required frequency of eye rolling one has to engage in, to be considered an eye roller?"[6] Adequate accounts of concepts must consider the range spanned by features. Feature matching fails this condition since it ignores that features span a possible range in the first place.

One reason for this difficulty is that the approach oversimplifies the relations between features and the objects that hold them. Sets of common and distinct features are modeled by countable whole numbers. This works in some cases, for instance, the seeds of an apple are countable and can be isolated from other features; an apple cannot have 2.5 seeds. However, this does not work in other cases. For instance, the colors of an apple are uncountable; apples typically have varying shades of green to yellow or red that transition smoothly.

Coming back to the case of perception, the assumption that the object-base, Δ, can be neatly decomposed into discrete features is difficult to combine with psychophysical models of perceptual-object representations, which commonly distinguish between 'inseparable' and 'separable' dimensions (Cheng & Pachella, 1984;

[5] Throughout this paper, I refer to concepts using small capitals.

[6] The example of eye-rolling illustrates that feature inference is at least partly determined by pragmatic factors. How much eye-rolling is needed to count as an eye-roller might depend on both the agent's experience with the frequency of the eye-rolling feature, and the agent's background assumptions concerning the possible variety of different cases of eye-rolling. Bayesian inference as elucidated below can capture such pragmatic effects reasonably well (cf. Qing et al., 2015), while the geometric and the feature-matching approaches fail to draw explicit distinctions between background knowledge and frequency of experience.

Gärdenfors, 2000; Melara, 1992). A set of dimensions is inseparable if an object that is assigned a value on one dimension must also be assigned a value along the other dimensions it is integrated with. For example, the dimensions hue, saturation and brightness are integral because it is impossible to represent a color shade along only the brightness dimension; it requires simultaneously assigning values along the hue and saturation dimensions. A set of dimensions is separable if it is possible to represent an object's property by assignment of values on a single dimension without assigning values along the other dimensions. For instance, an apple's shape can be represented independently of its color. This research suggests, contrary to the classical approach, that perceptual object representations are not always decomposable into discrete sets of features, making it difficult to explain the *perception* of similarities with feature-matching.

This claim is supported by various studies. Young children below the age of five typically confuse the height of a liquid in a container with the liquid's volume and just with time (typically after five) do they learn to distinguish between height and volume. However, at this stage, children can identify what color the liquid has.[7] This supports that decomposability of objects (e.g., liquids) into discrete features (e.g., height and volume) is not cognitively given and needs further explanation.[8] Furthermore, features associated with odors partly overlap with tastes and do not separate into discrete elements that form an exhaustive set (Jraissati & Deroy, 2021). In taking features as trivially decomposed, classical feature-matching ignores such evidence and suggestions that features themselves originate from similarities in perception.

A severe consequence of these issues is that classical feature-matching utilizes a too narrow notion of similarity that fails to generalize beyond the domain of judgement and voluntary decision-making. A broader notion is needed insofar as similarity representations exist in other domains, such as in perception and action. To see this limitation, recall that Tversky's explanation of directionality (Sect. 2.2) sets similarity judgements into a structural equivalence to similarity statements, which are grammatically analyzable and pertain to symbolic thought. This equivalence suggests that the contents of similarity judgements are Russellian propositions or Fregean senses. Following Tversky's analysis, when an agent judges how similar *a* is to *b*, the agent forms a representation with a content of the form "a is similar to b". This is distinct from its converse, "b is similar to a" since the contents of the two statements may differ in their truth values. Although Tversky (1977) makes no explicit mention of Fodor's (1975) work, his explanation of the relevant distinction comes close to ideas from the language of thought hypothesis. In this

[7] Initially, these studies were used to test Piaget's (1976, p. 177) theory of conservation, which hypothesizes that young children fail to understand conservation (e.g., conservation of volume when pouring a liquid from a wider into a narrower container).

[8] One explanation has been proposed by Gärdenfors (2000, p. 28), who argues that children learn to represent the different qualities of liquids by mentally dissociating distinct dimensions. While the perception of the world is inherently continuous—it does not contain intrinsically separate categories— children learn concepts by carving up the perceptual space into meaningful—discriminating—discrete and to be labelled chunks.

analogy, similarity comparisons take the form of linguistic propositions describing objects; features take the form of atomic symbols without internal structure. Can Tversky's account of similarity judgement be carried to the domain of perception? The difference between perceiving *a* as similar to *b* or vice versa may be attributed to differences in the accuracy conditions associated with these perceptions (cf. Siegel, 2010). Another motivation for attributing propositional content to perceptual similarity judgement is that such a position could explain how perceiving *a* and *b* as similar may justify certain inferences from beliefs about *a* to beliefs about *b*. In the conventional philosophical debate, conceptualists such as McDowell (1994) and Brewer (1999) have argued that perceptions provide reasons for belief only in the way beliefs justify other beliefs, and to satisfy this inferential aspect, perceptions must carry propositional content. However, the claim that the contents of perceptual similarity representations are propositional encounters a variety of problems.

Firstly, it is unclear why perceptual objects should stand in a grammatical relationship akin to the subject-object relationship in similarity statements. It is unclear how to understand grasp of these relations, since, in such cases, it is not trivial that the relevant structures are propositionally analyzable. According to proponents of perceptual nonconceptualism, perception is fundamentally nonpropositional and does not always require possession of concepts. For example, Peacocke's (1992) *scenario content* does not require the agent to possess the concept *C* to position the property *c* in a scenario. Heck (2000) argues that one can have a visual experience that represents that an object has a particular color shade without possessing the concept of that specific color shade. Following these views, perceptual similarity representations and similarity judgements remain distinct; possessing a concept is being in intentional states where contents are appropriately inferentially related, but the contents of perception are not so related and hence nonconceptual.[9] Additionally, philosophers of cognition argue that perceptual systems are informationally encapsulated and consist of 'modules' (Fodor, 1987), and thereby cannot compute over propositions in conceptual thought. Others appeal to a distinction in representational format between (nonpropositional) perception, analogue and continuous, and conceptual thought, which is propositional, digital, and discrete (Camp, 2007; Maley, 2011). These recent claims add severe doubts on the assumption that *perceptual* similarity is best analyzed in terms of propositional structure. Thus, insofar as classical feature-matching is limited by Tversky to the domain of judgement, it does not seem well suited as an account of perceptual similarity from the perspective of nonconceptualism.

Secondly, the focus on similarity statements raises the question whether children and animals who cannot identify grammatical or linguistic relationships (Glock, 2000) are capable to judge similarities and categorize the world. Within the psychology of animal learning and adaptive behavior, perceived similarity to a training stimulus often provides reasons for an animal to generalize a learned behavior to

[9] A popular argument for this position alludes to perceptual illusions, such as the Müller-Lyer illusion: even though I might strongly believe that the two lines have the same lengths, I continue to see them as having distinct lengths (see also Crane 1992, p. 150).

novel objects (Staddon, 2016). For instance, depending on how similar an apple and a pear look, an animal might be inclined to eat one upon eating the other. However, it is not trivial that these reasons must be bound to abilities of judgement and decision-making. Hurley (2003) argues convincingly that animal action builds on practical reason, outside the scope of "conceptualized inference or theorizing" (ibid., p. 231), suggesting that non-human animals and children may use similarity representations as a base for categorization, even if an appropriate explanation of why they categorize as they do may resist analysis in terms of similarity statements. Furthermore, Deroy (2019) discusses evidence in humans supporting the claim that at least perceptual categorization relies on involuntary mechanisms, not necessarily on abilities of conceptual thought and voluntary judgement.

Doubts on the exclusive focus on similarity judgements are further supported by recent perspectives of embodied approaches to cognition, which are rising in popularity and acceptance. These views challenge the traditional 'sandwich' model of information-processing systems (Hurley, 2002), where cognition forms the content, perception the input and action the output of the system. In contrast, proponents of embodied cognition argue that perception, cognition, and action are inherently intertwined. If the embodied-cognition view is correct, then most similarity representations (even the amodal ones) and the categorizations that build on them should be grounded in perception and sensorimotor activity (Barsalou, 2008; Harnad, 1990). From this perspective, the analysis in terms of similarity judgement is too narrow to capture the importance of similarity in perception, action, and cognition. It lacks an inclusive notion of similarity representation according to which perceptual categorization is embodied and action oriented. On the classical approach, features are discrete, but perception is continuous; features are extracted from a database of abstract objects, but they should be grounded in modal qualities.

As this debate illustrates, it cannot simply be assumed that perceptual similarity representations resemble conceptual thoughts in many ways, and so Tversky's exclusive focus on propositional, discrete, and abstract similarity judgement is by default too limited to develop a general account of similarity in cognition.

3.3 The Problem of Unprincipled Tweaking

The third problem with classical feature-matching is that it is too flexible. Tversky's contrast model accommodates various effects associated with similarity judgement (Sects. 2.2 and 2.3) but lacks an explanation for why these effects occur. The reasoning is that these cases arise from the unequal weighting and salience of distinct features.

The first problem with this line is that it likely results in a regress of additional terminologies (e.g., diagnosticity, salience, etc.) that call for further explanation. Take attribute salience as an example. Although salience is central to the feature-matching model, we still lack a detailed account of how it should be understood. Tversky explains directionality effects by arguing that more prominent objects (e.g., New York) are more salient, and therefore obtain higher weight in the contrast model than less prominent ones (e.g., Tel Aviv). This raises further questions: why

are more prominent objects more salient than less prominent ones? In what way are they more salient? How is salience measured?

In an initial attempt to explain the notion, Tversky explains that "[t]he factors that contribute to the salience of a stimulus include intensity, frequency, familiarity, good form, and informational content" (Tversky, 1977, p. 332), where intensity refers to an increase in the "signal-to-noise ratio, such as the brightness of a light, the loudness of a tone, the saturation of a color, the size of a letter, the frequency of an item, the clarity of a picture, or the vividness of an image". Diagnosticity, as illustrated earlier, refers to "the classificatory significance of features, that is, the importance or prevalence of the classifications that are based on these features" (Tversky, 1977, p. 342). Each of these answers invites further questions on the origin of attribute salience that remain unaddressed. In how far does salience relate to perceptual or attentional mechanisms or abstract reasoning skills? Is it (still) a core aspect of similarity judgement, or object recognition instead? If attribute salience is an aspect of perception and attention in a sense independent of prior categorizations and conceptual knowledge concerning the available objects, this is unsatisfying, since some account must be given of what determines changes in perceptual salience. Although feature matching is ubiquitous in computational psychology, detailing an account of attribute salience remains a problem to be solved.

The worry illustrated by attribute salience is that the approach lacks explanatory power. By continuing to add parameters to the model, proponents of feature-matching make their model unnecessarily complex. The model can be used to flexibly accommodate various specific findings associated with similarity and categorization, but this is only achieved by adding parameters (e.g., via the addition of feature weights and a salience scale) that can subsequently be tweaked in rather unprincipled ways. However, these additions are ad hoc. The associated flexibility does not show that the desired effects so accommodated come out as a core aspect of the feature-matching process, as opposed to other cognitive processes. Consequently, the model does not justify saying that directionality is a necessary feature associated with similarity judgement and might have to do with other aspects of attention or perception. The mere adjustment of the model parameters does not by itself explain what determines the parameters in the first place; it does not predict when and why the desired effects occur as a consequence of core aspects of the model.

3.4 Contrast to Geometric Models

A consequence of these three challenges is that Tversky's initial charges against approaches equivocating similarity with geometric distance (Gärdenfors, 2000; Shepard, 1962) appear to be unjustified, suggesting that the approach lacks some of the important advantages it initially suggested to have (most famously its ability to account for directionality and context-sensitivity effects). By definition, the geometric distance between two points, a and b, is the same as vice versa, suggesting that similarity is non-directional and independent of the order in which the two objects are compared. However, it is compatible with geometric models to explain directionality as an effect of additional cognitive processes that are sensitive to the

pragmatics associated with similarity statements, while not pertaining to aspects of similarity representations themselves.

This possibility is supported by Nosofsky (1986), who ascribes directionality effects to the influence of bias on perception, showing that the geometric model can accommodate directionality by adding a weight on dimensions that is formally similar to the weighting of features in Tversky's approach. However, Nosofsky (1986, pp. 54–55) explicitly dissociates "similarity representations", which remain subject to metric constraints, from additional "rather complex attention and decision processes" that operate on them. Similarity remains a metric relation between two objects while bias is "a characteristic pertaining to an individual object" (Nosofsky, 1991, p. 94). Thus, as an additional process, directionality is no case in point against using geometric models to explain judgements of similarity.

Furthermore, geometric models of similarity and categorization fare significantly better in accommodating the structure of perceptual similarity representations. Geometric distances can have a continuous structure and efforts have been made to accommodate the modality-specific nature of perceptual representations within this framework (Balkenius & Gärdenfors, 2016; Gärdenfors, 2007). Geometric models also have made progress on the problem of feature selection. For instance, Douven and Gärdenfors (2020) develop a set of design-principles and show how these can be instantiated under the assumption of a geometric-spaces model of similarity.

Taken together, these problems raise doubts that classical feature-matching advances discoveries of the mechanisms underlying similarity and categorization. In what remains, I elaborate three constraints to address these challenges and positively develop the approach given recent developments in computational cognitive science. I subsequently discuss novel insights that can be gained from this perspective for grounding cognition in similarity as inference.

4 Similarity as Bayesian inference

My response to these challenges is that feature matching should be seen as a kind of Bayesian inference. Bayesian models of cognition have been proposed to characterize mechanisms of perception (Knill & Richards, 1996) and motor learning (Körding & Wolpert, 2004), conceptual thought (Rescorla, 2009; Tenenbaum, 1999), and many more diverse cognitive capacities. The various applications of Bayesian modeling in the literature illustrate that this framework is suited to model a variety of inference tasks, including perception and categorization. I propose that this framework is therefore well suited to improve upon Tversky's approach; it offers a general computational account of feature-matching[10] that is not limited to Tversky's

[10] The account of similarity as Bayesian inference classifies as computational since it remains neutral on how exactly feature representations must be individuated in the agent's mind; no causal mechanisms are cited in explaining how features are inferred by the brain or body. Although the initial definition of a computational account as it appears in Marr's (1982) celebrated framework is by now outdated (there seem to be many more kinds of levels than initially proposed between the computational, algorithmic, and implementational levels, see McClamrock 1991, Danks 2008, and Hardcastle & Hardcastle 2015), computational analyses remain important scientific tools for the discovery and selection of mechanistic models (Colombo & Hartmann, 2017; Zednik & Jäkel, 2016).

preferred narrow class of similarity judgements. Indeed, viewing similarity as Bayesian inference provides the relevant foundation of Tversky's approach by appeal to probabilistic principles that obtain regardless of how one glosses over feature representations (e.g., as geometric distances or cardinalities of sets). This novel view thus promises to be general enough to capture similarity judgements in a variety of domains, such as monetary or aesthetic value, but also colors or odors. On this view, three inductive biases act as constraints on the inference of a feature or dimension.

4.1 The Contrast Class as an Internal Constraint on Similarity

The first constraint is the consideration of the contrast class as a background of a similarity comparison. Following Rosch (1978), a contrast class is the set of items that occur at the same level of organization in a classificatory taxonomy. For example, HORSE and PIG are contrast classes for DEER; they occur at the same level of inclusiveness of the category of ungulates. The contrast class constrains the range of a class by defining its borders to other classes. Adopting this idea, Bloch-Mullins (2021) argues that the range spanned by the items in a contrast class is essential to understand the nature of similarity. The key to her account is the addition of a 'foil' as a constitutive constraint on the distribution of values associated with a feature. The foil is the collection of the items in the background against which targets are compared, it is not limited to the particulars one is currently exposed to but forms part of one's background knowledge.

To borrow an example from Bloch-Mullins (ibid., p. 39), the statement "the cost of living in Seattle is similar to the cost of living in New York City" is true with regards to all cities in the world, but false with regards to only US cities. The choice of the relevant contrast class determines the difference, which occurs because Seattle and NYC are aligned more closely than most other cities in the world along the dimension 'cost of living', but where they are not closer than most cities in the US. Generally, similarity is structured such that one initially forms a range of acceptable values along a given dimension and subsequently infers to what extent the properties of the target objects fall within that range. The range of attributes shared by the contrast class (i.e., its dispersion) defines the range of a feature along a dimension; this forms an integral part of similarity, as opposed to external processes of inference and decision. Thus, an important aspect of the foil account is that it turns similarity into a three-place predicate. Not only is it relevant to similarity what value a feature obtains, but also how the values of features correlate.

Although Bloch-Mullins (2021) does not explicitly draw this connection, her account sits well within core ideas of Bayesian principles of reasoning (Kemp, Bernstein, & Tenenbaum, 2005; Navarro & Perfors, 2010; Tenenbaum & Griffiths, 2001). In the following, I explore this connection to advance classical feature matching. The hallmark of these models is that they explain why similarity takes the form that it does, as opposed to focusing only on questions concerning how similarity is computed, as is typical in classical feature-matching. This makes them ideally suited to address the problems discussed earlier.

4.2 Similarity as an Inductive Inference Task

Bayesian models of cognition view similarity as an inductive inference task. That is, as a task of inferring why a certain set of features that are given as inputs should be inferred as relevant for a similarity-judgement or categorization task. Often, the logic of its solution is identified in terms of intuitive principles of rationality (e.g., information-gain or survival, cognitive economy).

4.2.1 The Size Principle for Feature-Discovery

A promising proposal along these lines is the size principle for feature discovery (Navarro & Perfors, 2010; Tenenbaum & Griffiths, 2001). It states that the likelihood of a feature, F, corresponds to the ratio of the number of instances that possess F (e.g., the number of things that eat peanuts for F: 'eats peanuts'). The size principle predicts preferences for rare (i.e., smaller) features—when two objects match on a rare feature, they are relatively more similar than when they match on a ubiquitous feature.

The size principle responds to the feature-selection problem insofar as F is selected based on its size, which depends on its frequency. The rationality associated with the size principle is to select highly informative features. For example, when two objects (e.g., a tiger and a bird) mismatch on a rare feature (e.g., feathered), they are relatively more dissimilar than when they mismatch on a ubiquitous feature (e.g., two-legged).

However, although Navarro and Perfors advertise their approach as broadly Bayesian, they do not specify constraints on prior probabilities. Bayes' theorem combines priors and likelihoods according to the following scheme:

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)},$$

where the likelihood, $Pr(E|H)$, indicates the probability of observing E if the hypothesis H was true. This is multiplied with the prior, $Pr(H)$, which indicates the probability of H regardless of E, to provide the posterior probability of H given E. $Pr(E)$ is a normalization term. The size principle specifies the likelihood term indicating how probable it is to observe E if it had F and it does not specify the priors. However, considering the size principle alone is often insufficient and counter-intuitive. When we have the candidate concepts EDIBLE and EDIBLE OR POISONOUS, it is more adaptively successful for an agent who does not know the true concept to assume that a given instance falls under the larger, disjunctive, concept. In this case it is reasonable to choose the larger concept, even though the evidence is logically compatible with both concepts. Such cases seem to be influenced by prior beliefs about how features are distributed, regardless of the available evidence.

4.2.2 Variability-Based Priors

To fill this gap, I suggest furnishing the probabilistic approach with a third ingredient: the variability-based diagnosticity principle proposed by Goldstone et al. (1997), according to which high-variability features attain higher probability or weight. Variability is defined in terms of the range of different values a feature takes. The key is that the range spanned by a feature matters in determining its diagnostic relevance to a categorization task—a feature is more diagnostically relevant if it has a high variability or spans a broad range of possible values. High-variability features carry more information than low-variability features in the sense that "dimensions that have many different values will also have a greater degree of difference between dimension values than will dimensions that have fewer different values" (Goldstone et al., 1997, p. 243). This may allow agents to better distinguish objects that fall somewhere along these dimensions.

Consider the feature TIME OF DAY, which is diagnostically relevant for the comparison between sunrise and sunset because they are opposites along this dimension, that is, it is a high-variability feature (Medin et al., 1993). Conversely, low-variability features are irrelevant, i.e., dimensions along which a set of alternatives do not differ greatly (e.g., WARM COLOUR) will be ignored. In this sense, the variability criterion for feature selection offers a first step to narrowing down the items in the background of a similarity-judgement task (e.g., the varying times of the day, as opposed to the range of warm colors). Goldstone and colleagues found evidence for the variability-based diagnosticity principle in an experiment that asked participants to choose one of two alternative objects that best matched a standard object. They compared a 'shared match' condition, in which two objects shared a feature that matches the prototype, with a 'shared mismatch' condition, in which the objects shared features that mismatched the prototype. They predicted that the shared-mismatch condition contains a greater variability on the relevant feature, and so subjects should judge the similarity of one of the objects to the standard to be greater. The results of their study confirmed this; subjects preferred choosing the object that matched along high-variability dimensions.

Unfortunately, Goldstone et al.'s principle has received little attention in the Bayesian literature, where feature variability can be understood in terms of the mean and variance of a probability distribution that encodes the subjects' prior predictions about the possible range of a feature. The prior preference for high-variability features can be combined with a preference for less frequent features, following the size principle. For instance, the prior probability of the feature `crimson' may initially be lower than for `red' because, intuitively, its instances vary less along the hue dimension. However, although the prior probability is, intuitively, higher for `red', `crimson' better predicts evidence that is compatible with both features, and so it obtains a higher likelihood. If features should be selected according to their posterior probability, which optimally combines these two constraints under Bayes' theorem, then feature specification underlies a compromise between both the size and range of a feature.

There are interesting parallels of this Bayesian account to Carnap's (1980) posthumously published work on inductive logic and its recent expansion in cognitive

science by Decock et al. (2016) and Poth (2019). These approaches derive the agent's prior degrees of belief (e.g., about what feature is relevant in an inductive inference task) from the geometric structure of concepts. Specifically, Decock et al. (2016) propose a *geometric principle of indifference* according to which the a priori probability for an object, o, to fall under a concept C is proportional to the *size* of the region that stands for the concept in a conceptual space. Following Carnap, the prior probability of the concept corresponds to the area of its region relative to conceptual space. This result coincides with the preference for variability-based priors in Goldstone et al.'s work, where areas covering a broader range of values along a dimension obtain higher prior probability. Despite lack of empirical knowledge, the agent still has reason to treat regions with different sizes as differently plausible candidates in inferences of properties (for discussion, see Poth, 2019).

Readers might object that Bayesian models miss out on directionality. However, the Bayesian approach fails directionality only in the sense in which Tversky's contrast model is directional. Indeed, it is directional in another sense since the conditional probability functions (1) $pr(A|B) = pr(A \cap B)/pr(B)$ and (2) $pr(B|A) = pr(A \cap B)/pr(A)$ likely obtain different results. Assume A represents the hypothesis that an object is crimson red, and B is the hypothesis that the object is edible. The outcome in (1) is likely to be different from (2) because (1) depends only on whether the object is edible, and (2) depends only on whether the object is red. But the probability of a random object being red seems to be very much independent, and hence likely to be unequal to the probability of it being edible. This sort of directionality might deliver a better explanation of similarity, since it comes out as a core aspect of the Bayesian model (i.e., the definition of conditional probability), while the contrast model only manually accommodates it with parameters that were introduced solely for this purpose (Poth 2022).

4.3 Responding to the Challenges

This novel understanding of feature-matching at least partially responds to the three problems discussed in Sect. 3. Firstly, it addresses the feature-selection problem as one of inductive inference of the feature with the highest posterior probability. It is assumed that likelihoods are constrained by the size principle or the frequency at which features occur, and priors are constrained by the expected variability of a feature. By optimally combining these constraints in the manner of Bayes' theorem, agents infer maximally informative features. That is, features (defined conditional on a contrast class) should be selected according to their posterior probability, which trades high-variability priors (favoring broad features) with the size principle (favoring narrow features) under Bayes' theorem.

Secondly, the probabilistic approach offers a more plausible view of feature representations, which are not binary and discrete but constituted by a range of values as determined by the distributions of items in the foil. This brings feature-matching closer to the inherently continuous nature of perception and conceptual thought. Insofar as relations between features are gradual (e.g., the cost of living

in New York is *more or less* like the cost of living in Seattle given US cities), feature representations vary gradually according to the strength of inference. Bayesian perceptual psychology sees no structural difference between perception and belief, as both have probabilistic structures. It highlights their gradual nature by replacing the conventional notion of outright belief with the notion of degree of belief. Proponents posit that all aspects of cognition, including perception, take the form of probabilistic guesses or expectations about how the world unfolds (Clark, 2013; Hohwy, 2013). But they do not take this to mean that perception requires concepts or that it has propositional content as classically construed. Perceptual states represent, not that there is a green shade, but a probability distribution over multiple green shades (cf. Sprevak, 2020), and a perceptual system in such a state may then `decide' on a single shade based on the highest posterior distribution by maximizing expected utility. In perception, probabilistic beliefs operate subpersonally; they are implicit representational states and unavailable to explicit judgement. They are hence perfectly suited to provide a notion of implicit, perceptual, similarity judgement. In this sense, Bayesian modeling allows for a broader conception of similarity than classical feature-matching and captures at least some of the important aspects associated with perception and action. Whereas Tversky initially explains that the context determines the feature weights, probabilistic approaches add to this, more specifically, that the weights are determined by the subjective probability distribution over observed features. The magnitude of a feature might depend on how often it is activated, and perceptual systems might have subpersonal-level expectations about how frequent a feature is and update this expectation based on novel experiences in the manner of Bayesian inference. Generally, modeling cognition as a form of Bayesian inference has proven fruitful to model aspects of perception and action in this sense (e.g., Körding & Wolpert, 2004). Aside from advances in computational neuroscience, recent work in philosophy suggests that Bayesian learners may infer similarities without the ability for linguistically-analyzable thought. For instance, Rescorla (2009) shows that Bayesian reasoning can be performed with probability distributions structured by geometric relations, which lack predicate-argument structure and compositionality. He illustrates this with the case of a robot, which updates its beliefs about its position relative to several landmarks based on its previous motor commands and sensory inputs. Generally, the approach remains neutral on whether the size of a feature should be defined in terms of geometric distances, consequential subsets in an abstract stimulus space, or in a language of thought structure (Tenenbaum & Griffiths, 2001; Tenenbaum et al., 2011). Thus, rather than dividing similarity in perception from similarity judgement, Bayesian modeling establishes similarity as a unified property of inference in perception, action, as well as conceptual thought. These considerations provide a partial response to the initial problem of cognitive plausibility, as viewing similarity as a form of Bayesian inference seems to capture more aspects associated with cognitive systems than classical feature-matching.

Nevertheless, this response is only partial—the cognitive plausibility of Bayesian computation has been questioned on various grounds, including its computational intractability (van Rooij et al. 2018; Kwisthout & van Rooij, 2020). Realism about

Bayesian cognition is not mainstream (but see Rescorla, 2019) and there are serious challenges to taking Bayesian computation literally, given that the empirical evidence is often unequivocal and considering available alternatives (Colombo et al., 2020). A full discussion of the appropriateness of Bayesian-inference explanations of cognition is outside the scope of this paper, but this ongoing debate motivates further investigation of the overall cognitive plausibility and instrumental value of viewing similarity as Bayesian inference, which remains an open challenge.

Thirdly, the novel approach bears explanatory advantages in drawing out similarity as a core consequence of probabilistic reasoning principles at the computational level. While classical feature-matching focuses on questions concerning how similarity is computed, little time is spent on asking why people judge similarity in the way they do, or why they should. The Bayesian approach furnishes the diagnosticity principle (Sect. 2.2) with a theoretically refined understanding of similarity as an inference, under which certain ways of computing similarity can be discarded given the available rationality constraints (Zednik & Jäkel, 2016). This is illustrated with the size principle, where rarity provides a rational reason to select a feature as relevant in a similarity-judgment task, hence combatting unprincipled tweaking. In other words, the effects of similarity judgement come out directly from principles of probabilistic inference.Finally, these Bayesian principles give a computational explanation of feature-matching that aligns with recent developments in computational cognitive science. Following van Rooij and Baggio (2021), developing good theories in psychology requires iterative progressing through a theoretical cycle, whereby an initial theory is iteratively revised and refined. Along these lines stands my analysis of similarity as an inductive inference task. The intuitive theory is that people judge similarities, which Tversky explicates formally with the contrast model. In analyzing the theory, I have checked for conceptual errors in Tversky's approach in Sect. 3, and subsequently refined it in the context of Bayesian models in Sect. 4, where I have introduced a set of theoretical constraints. Following van Rooij and Baggio, this step can help to "narrow down the space of possible functions to those describing real-world capacities" (van Rooij & Baggio, 2021, p. 690). The size principle and variability-based priors are steps in this direction. They narrow down the set of all possible features that could be inferred (e.g., based on their logical compatibility with a hypothesis) to a subset of informative features. What is also highlighted by van Rooij and Baggio is the need for a computational-level theory (Marr, 1982) when analyzing psychological capacities. The Bayesian approach improves classical feature-matching on this front in that it unifies similarity and categorization under a single computational-level description of inductive inference. In the end, a theoretical cycle involves iterations, of which I have only presented the first step. Further constraints may improve this description of similarity. For instance, an outstanding area of advancement is to show how this analysis can be integrated with an embodied cognition perspective.

5 Implications for Similarity-Based Accounts of Cognition

A central consequence of this novel view of similarity is that feature representation is cognitively derivative of probabilistic inference over geometric information. Contrary to the common lore in psychology, similarity is not theoretically fundamental in explaining generalization and inference. The two major kinds of similarity considered in this debate are integrated with a Bayesian approach that outsources geometric properties associated with perceptual space to infer the most relevant feature based on a (non-exhaustive) set of rationality constraints. This approach not only explains (a) feature-selection, (b) cognitive plausibility and (c) the core principles based on which similarity is modulated in context. It also highlights the underappreciated insights from Tversky, that similarity representations are context sensitive, e.g., they depend on how the range of values along one dimension (e.g., `location') influences the distribution of values along another (e.g., `cost of living'). The probabilistic framework is optimally suited to accommodate this aspect in terms of mutual-informational relevance relations among perceptual dimensions. Knowing that my friend lives in New York makes them more likely to pay high living costs, while knowing that the temperature in New York is high is irrelevant—temperature changes bear no significant effect on changes in living costs. What integrates the two models of similarity, contrary to their classical opposition, is that probabilistic inference explains similarity either way as an outcome of core principles of Bayesian inference.

Contrary to common lore in philosophy, Tversky's conceptual similarity does not necessarily originate in higher-level cognitive processes but in basic probabilistic computations over sets of features in perceptual space. Similarity as Bayesian inference provides a broader framework that remains neutral on whether the capacity to perceive similarity requires the capacity for conceptual thought. It also offers an opportunity to connect the feature-matching and geometric approaches by grounding them both in principles of statistical inference. One possibility is that those geometric properties associated with perceptual dimensions justify the inference of feature overlap. Brössel (2017) has recently applied a similar idea using the theory of conceptual spaces (Gärdenfors, 2000) to structurally relate color perception and perceptual beliefs about color. He assumes that perceptual color experiences with non-conceptual content are points, and color concepts are regions in perceptual similarity space. He establishes the rational relationship between the content of color experiences and the conceptual content of beliefs about color by introducing a credence function that is defined over both, relations among color experience and color concepts, in this space. While color concepts are ``understood as binary random variables that take on the value 1 if the relevant object's shade of color falls under the given color concept and 0 otherwise'' (ibid., p. 735), color experiences are understood as continuous random variables associated with the dimensions of color perception. Brössel's probabilistic inference account thus illustrates how judgements about color can be probabilistically related to color perception,[11] and so judging the

[11] Although an application to the case of odor remains to be established in future work, a recent analysis offered by Jraissati & Deroy (2021) is a promising first step. According to the authors, what seems to be crucial to odor concepts is that they are local; they do not occupy the space of possible odor percep-

similarity between color shades fits very well within a probabilistic account of similarity as an inference.

Note that there is nothing special about context and directionality effects in probabilistic reasoning about similarity. These effects may likewise influence probabilistic computation at the subpersonal level of perceptual thought. For instance, my implicit beliefs may influence how similar I perceive two paintings, but this need not be an influence of higher-level cognition. In the classroom, the original Mona Lisa and a copy of it might be represented as more similar to each other than each of them to a Magritte, based on their (inferred) perceptual features. In the context of an art auction, an expert's implicit knowledge of the price of the original Mona Lisa and the original Magritte might make them more mutually similar than the original Mona Lisa and its copy. From the probabilistic point of view, such implicit background beliefs modulate feature selection and similarity computation across multiple distinctions of cognition (e.g., across the divide between perception and conceptual thought, and the divide between implicit and explicit belief).

6 Conclusion

In this paper, I have discussed three key problems for the feature-matching approach to similarity and categorization, which is widely used in cognitive science but rarely explicitly justified considering recent discussions on the nature of cognition. In revisiting the approach, I have suggested combining three new ingredients: consideration of the contrast class as a constitutive constraint on similarity, the size principle, which selects for rare features, and variability-based priors, which impose a preference for high-variability features. I have embedded these ingredients in a Bayesian approach to cognition that offers promising, albeit incomplete, ways of meeting the given challenges in terms of a computational analysis of similarity as an inductive-inference task. An advantage of this novel perspective is that it avoids a view on perceptual similarity as pertaining to conceptual thought while maintaining rational relations to similarity judgement. With these adjustments, feature-matching might be further developed as a viable option to explain how psychological similarity serves as a viable ground for cognition across the divide between perception and conceptual thought.

Acknowledgements I am indebted to Mark Sprevak, Alistair Isaac, Peter Brössel and Insa Lawler for invaluable feedback on earlier drafts of this paper. I am grateful to the audience at the joint ESPP/SPP 2022 conference at the University of Milan and two anonymous reviewers of this journal for their insightful comments to improve this work.

Footnote 11 (continued)

tions in a mutually exhaustive and finite way. Although Bayesian inferences are classically defined over Boolean algebras, it is not a sign of irrationality if perceptual odor concepts fail to exhaust the space of possibilities, since such incompleteness is not incoherent with the axioms of probability. It remains a possibility that probabilistic inferences could be defined over incomplete spaces of perceptual odor concepts.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by a research fellowship from Ruhr-Universität Bochum.

References

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

Balkenius, C., & Gärdenfors, P. (2016). Spaces in the brain: From neurons to meanings. *Frontiers in Psychology, 7*, 1820.

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

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

Bloch-Mullins, C. L. (2020). Bridging the gap between similarity and causality: An integrated approach to concepts. *The British Journal for the Philosophy of Science,* 69(3).

Bloch-Mullins, C. L. (2021). Similarity reimagined (with implications for a theory of concepts). *Theoria, 87*(1), 31–68.

Brewer, B. (1999). *Perception and Reason*. Oxford University Press.

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

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives, 21*, 145–182.

Carnap, R. (1980). A basic system of inductive logic part ii. *Studies in Inductive Logic and Probability, 2*, 7.

Cheng, P. W., & Pachella, R. G. (1984). A psychophysical approach to dimensional separability. *Cognitive Psychology, 16*(3), 279–304.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Colombo, M., Elkin, L., & Hartmann, S. (2020). Being realist about Bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science, 72(1).*

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science, 68*, 451–484.

Crane, T. (1992). The nonconceptual content of experience. In T. Crane (Ed.), *The contents of experience: Essays on perception* (pp. 1–22). Cambridge University Press.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 59–75).

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

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

Deroy, O. (2019). Categorising without concepts. *Review of Philosophy and Psychology, 10*(3), 465–478.

Douven, I., & Gärdenfors, P. (2020). What are natural concepts? *A Design Perspective. Mind & Language, 35*(3), 313–334.

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017). Measuring graded membership: The case of color. *Cognitive Science, 41*(3), 686–722.

Falkowski, A., Sidoruk-Błach, M., Bartosiewicz, Z., & Olszewska, J. M. (2018). Asymmetry in similarity formation: Extension of similarity theory to open sets of features. *The American Journal of Psychology, 131*(2), 151–159.
Fodor, J. A. (1975). *The Language of Thought* (Vol. 5). Harvard university press.
Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge: MIT Press.
Galesic, M., Goode, A. W., Wallsten, T. S., & Norman, K. L. (2018). Using Tversky's contrast model to investigate how features of similarity affect judgments of likelihood. *Judgment & Decision Making, 13*(2), 163–169.
Gärdenfors, P. (2007). *Cognitive semantics and image schemas with embodied forces*, In Krois, J.M., Westerkamp, D., Steidele, A., Rosengren, M. Embodiment in Cognition and Culture, John Benjamins Publishing Company, pp 57–76.
Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.
Glock, H.-J. (2000). Animals, thoughts and concepts. *Synthese, 123*(1), 35–64.
Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition, 65*(2–3), 231–262.
Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition, 25*(2), 237–255.
Goodman, N. (1972). *Seven strictures on similarity. In Problems and projects (1st* (print). Bobbs-Merrill.
Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape- similarity judgments. *Psychological Science, 20*(4), 447–454.
Hardcastle, V. G., & Hardcastle, K. (2015). Marr's levels revisited: Understanding how brains break. *Topics in Cognitive Science, 7*(2), 259–273.
Harnad, S. (1990). The symbol grounding problem. *Physica d: Nonlinear Phenomena, 42*(1–3), 335–346.
Heck, R. G. (2000). Nonconceptual content and the "space of reasons." *Philosophical Review, 109*(4), 483–523.
Hohwy, J. (2013). The predictive mind. Oxford University Press. Hurley, S. (2002). *Consciousness in action*. Cambridge: Harvard University Press.
Hurley, S. (2002). *Consciousness in action*. Harvard University Press.
Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language, 18*(3), 231–257.
Jraissati, Y., & Deroy, O. (2021). Categorizing smells: A localist approach. *Cognitive Science, 45*(1), e12930.
Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1132–1137).
Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.
Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), 244–247.
Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology, 12*(1), 4–34.
Krawczak, M., & Szkatuła, G., et al. (2018). On asymmetric problems of objects' comparison. In L. Rutkowski (Ed.), *Artificial intelligence and soft computing* (pp. 398–407). Springer International Publishing.
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.
Kwisthout, J., & Van Rooij, I. (2020). Computational resource demands of a predictive Bayesian brain. *Computational Brain & Behavior, 3*(2), 174–188.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X16001837
Langkau, J., & Nimtz, C. (2010). *New perspectives on concepts* (Vol. 81). Rodopi.
Machery, E. (2007). Concept empiricism: A methodological critique. *Cognition, 104*(1), 19–46.
Maley, C. J. (2011). Analog and digital, continuous and discrete. *Philosophical Studies, 155*(1), 117–131.
Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman.
McClamrock,. (1991). R. Marr's three levels: A re-evaluation. *Minds and Machines, 1*, 185–196. https://doi.org/10.1007/BF00361036
McDowell, J. (1994). *Mind and World*. Harvard University Press.
Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254.

Melara, R. D. (1992). *The concept of perceptual similarity: From psychophysics to cognitive psychology*. In Advances in psychology, pp 303–388.
Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica, 133*(3), 256–268.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39.
Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23*(1), 94–140.
Peacocke, C. (1992). *A study of concepts*. The MIT Press.
Piaget, J. (1976). Identity and conservation. In B. Inhelder, H. H. Chipman, & C. Zwingmann (Eds.), *Piaget and his school: A reader in developmental psychology* (pp. 89–99). Berlin Heidelberg: Berlin, Heidelberg Springer. https://doi.org/10.1007/978-3-642-46323-5_8
Poth, N. L. (2019). Generalisation probabilities and perceptual categorisation. In M. Kaipainen, F. Zenker, A. Hautamäki, & P. Gärdenfors (Eds.), *Conceptual spaces: Elaborations and applications conceptual spaces* (pp. 7–28). Cham: Springer.
Poth, N. (2022). Refining the Bayesian approach to unifying generalisation. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-022-00613-5
Qing, C., & Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In H. Zeevat & H.-C. Schmitz (Eds.), *Bayesian natural language semantics and pragmatics* (pp. 201–220). Cham: Springer.
Rahnama, J., & Hüllermeier, E., et al. (2020). Learning Tversky similarity. In M.-J. Lesot (Ed.), *Information processing and management of uncertainty in knowledge-based systems* (pp. 269–280). Springer International Publishing.
Rescorla, M. (2019). A realist perspective on Bayesian cognitive science, In Inference and consciousness, Anders Nes & Timothy Chan eds Routledge, pp 40–73.
Rescorla, M. (2009). Cognitive maps and the language of thought. *The British Journal for the Philosophy of Science, 60*(2), 377–407. https://doi.org/10.1093/bjps/axp012
Restle, F. (1961). *Psychology of judgment and choice: A theoretical essay*. Springer: Wiley.
Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*(6), 665–681.
Rosch, E. (1978). *Principles of categorization*. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorization. New Jersey: Lawrence Erlbaum Associates.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605.
Sanborn, A. N., Heller, K., Austerweil, J. L., & Chater, N. (2021). Refresh: A new approach to modeling dimensional biases in perceptual similarity and categorization. *Psychological Review, 128*(6), 1145.
Shanon, B. (1988). On the similarity of features. *New Ideas in Psychology, 6*(3), 307–321.
Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function i. *Psychometrika, 27*(2), 125–140.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.
Siegel, S. (2010). *Do experiences have contents?* In Bence -Nanay (ed.), Perceiving the World, Oxford University Press
Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science, 360*(6389), 652–656.
Sjöberg, L. (1972). A cognitive theory of similarity. *Goteborg Psychological Reports*, 2(10).
Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition, 65*(2–3), 87–101.
Sprevak, M. (2020). Two kinds of information processing in cognition. *Review of Philosophy and Psychology, 11*(3), 591–611.
Staddon, J. E. R. (2016). *Adaptive behavior and learning*. Cambridge University Press.
Tenenbaum, J. B. (1999). A Bayesian Framework for Concept Learning (Doctoral dissertation, Massachusetts Institute of Technology).
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629–640.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (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.
Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and Categorization, 1*, 79–98.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high- verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682–697.
van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese, 195*(2), 491–510.
Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese, 193*(12), 3951–3985.