Conceptual Spaces Between Mind and World: Naturalness, Interaction, and Explanation

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philosophie (Dr. phil.)

durch die Philosophische Fakultät der

Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Sebastian Scholz

aus

DORTMUND

Betreuer/in: Prof. Dr. Gottfried Vosgerau

Düsseldorf, Oktober 2025

D61

Acknowledgements

To my dear friends and family, supportive supervisors, and splendid peers with whom I could share a thought after some lecture or over a balmy evening: Thank you for providing an environment in which I could learn and grow. I deeply appreciate you.

Special thanks to my supervisor Gottfried Vosgerau, whose continued input and support have made this dissertation possible in the first place. Thanks to my supervisors Jan G. Michel and Gerhard Schurz. Thanks to Igor Douven, Matías Osta-Vélez, and Nina Poth for helpful discussions.

Thanks to my friends and colleagues at Heinrich Heine University Düsseldorf for their feedback and support: René Baston, Daian Bica, Natalie Böddicker, Jonas Carstens, Alexander Christian, Lukas Gallach, Giacomo Giannini, Tim Grasshöfer, Susanne Hahn, Laura Hartmann-Wackers, Paul Hasselkuß, David Hommen, Sara Ipakchi, Eva Kellner, Benedikt Kenyah-Damptey, Felix Kopecky, Chiara Lisciandra, David Löwenstein, Julia Mirkin, Nastasia Müller, Julian Pöhling, Katarzyna Rajska-Kaselow, Frenzis Scheffels, Markus Schrenk, Marion Seiche, Maria Sekatskaya, Maria Sojka, Dennis Sölch, Oliver Victor, Simon Wimmer.

Thanks to my friends and colleagues at Lund University for an unforgettable research – and hiking – visit: Dani Alnashi, Andrey Anikin, Thibault Boehly, Vidar Bratt, Niklas Dahl, Anton Emilsson, Sebastian Goksör, Hubert Hågemark, Martin Jönsson, Jiwon Kim, Pierre Klintefors, Sarah Köglsperger, Lona Lalic, Marianna Leventi, Elsa Magnell, Jenny Magnusson, Shervin Mirzaeighazi, August Olsen, Fredrik Österblom, Manuel Quezada, Simon Rosengren, Sophie Serra, Martin Sjöberg, Samantha Stedtler, Jakob Stenseke, Andreas Stephens, Trond Arild Tjøstheim, Melina Tsapos, Balder Ask Zaar. Special thanks to Peter Gärdenfors for making this possible and for taking the time to discuss these ideas with me.

Thanks to my dear friends for their support: Jonas Böhmann, Bijan Fink, Tomas Kleiner, Fiona Schrading, Christian Vogt.

Thanks to my parents, Anette Münker and Volker Scholz, as well as to my wonderful sister, Stefanie Scholz. Thanks also to Christel Scholz, who raised her family well despite hardship and loss.

Finally, thanks to Andrea Schneider for our shared time on this peculiar, fragile planet.

Contents

Part I – Introductory Paper 1
1. Introduction 1
2. Conceptual Spaces 2
    2.1. Core Ideas 3
    2.2. Concepts and Mental Representations 6
    2.3. Similarity 8
    2.4. Reasoning 9
3. Papers 11
    3.1. (A) Goodman's Riddle 11
        3.1.1. The Realism Debate 11
        3.1.2. Results 12
    3.2. (B) Cognitive Sparseness 12
    3.3. (C) Predictive Models 13
        3.3.1. Probability 14
        3.3.2. Results 15
    3.4. (D) Semantically Deficient Models of Cognition 16
4. Developments and Results 17
References 19

Part II – Publications .....
    Paper A: Conceptual Spaces: A Solution to Goodman's New Riddle of Induction? 25
    Paper B: Conceptual spaces: Naturalness or cognitive sparseness? 46
    Paper C: Conceptual spaces as predictive models 68
    Paper D: Semantically Deficient Models of Cognition 97

Part I – Introductory Paper

1. Introduction

This dissertation examines the data format of thought. Historically, rationalists and empiricists debated whether thinking is fundamentally linguistic or imagistic. Today, empirically informed philosophers of mind draw on mathematical frameworks from cognitive psychology, linguistics, and related disciplines that together constitute contemporary cognitive science. Peter Gärdenfors's theory of Conceptual Spaces (CS) is one such framework. It models thinking geometrically – a stance with rich philosophical implications that coalesce around the themes of naturalness, interaction, and explanation.

The preceding paragraph is an elevator pitch for a broad audience. As such, it involves several oversimplifications: The term "data format" evokes digital computation and suggests that CS concerns the *form* of thought, while it really concerns (non-digital) thought *content*. Moreover, "thought" is the wrong level of granularity because CS targets concepts – the building blocks of thought. Rationalism and empiricism are broad positions in epistemology, and are not automatically committed to specific views on the structure of thought contents. And debates over whether thinking is linguistic or imagistic remain very much alive. What follows, therefore, is a more elaborate but still reader-friendly introduction for philosophers and cognitive scientists who are unfamiliar with the subject.

CS are geometric models which explain and predict cognitive behavior. As a geographic map represents the organization of a landscape, CS represent how individuals or groups organize semantic knowledge. Whereas distances on the map represent distances in the landscape, distances in CS encode similarity relations among possible objects or concepts. And whereas the axes of a map represent cardinal directions, the dimensions of CS represent any feature that objects may possess. Concepts are thought to be regions, or sets of possible objects in CS. Thus, CS is a theory about the structure of conceptual content. Concepts are considered "natural" when its region exhibits cohesive geometric properties: similar objects cluster, and cognitively economical agents group them accordingly.

Several versions of CS, or CS-adjacent theories exist in the literature. This dissertation, however, focuses on Peter Gärdenfors's account because it is well-developed and has several features that make it particularly interesting for a philosophical investigation. Most notably, the cognitive account of naturalness outlined above is distinctive to his approach. The papers

comprising this cumulative dissertation critically engage different aspects of Gärdenfors's theory:

(A) Scholz (2024): Conceptual Spaces: A Solution to Goodman's New Riddle of Induction?

(B) Scholz & Vosgerau (2025): Conceptual spaces: Naturalness or cognitive sparseness?

(C) Scholz (2026a): Conceptual Spaces as Predictive Models

(D) Scholz (2026b): Semantically Deficient Models of Cognition – A Bayesian Case Study, Conceptual Spaces, and AI Opacity

(A) and (B) investigate naturalness, (C) and (D) investigate interaction, and (D) investigates explanations. These themes, however, run through all four papers. Taken together, the dissertation delivers three core contributions. First, it clarifies naturalness by distinguishing cognitive sparseness from objectivity and by defending a methodological route to the latter grounded in scientific practice. Second, it advances an account of interaction by situating CS within predictive, Bayesian modeling. Third, it articulates an explanatory standard for cognitive science and shows how Bayesian models can meet this standard by adapting CS.

The papers also trace a trajectory along several dimensions of intellectual development: from realism to pragmatism, from metaphysics to cognition, and from philosophy of science to the philosophy of cognitive science. Articulating these developments is a primary aim of this introductory paper. A second aim is to introduce the terminology and philosophical backdrop needed to situate my contributions. A third aim is to present the results of the papers concisely.

I proceed as follows. Section 2 introduces the CS account in more detail and clarifies key concepts that recur throughout the dissertation. Section 3 presents the core ideas and accomplishments of the individual papers, together with key concepts specific to each. Section 4 reflects on the intellectual developments and results which emerge from the dissertation as a whole.

2. Conceptual Spaces

This section introduces the CS framework by outlining its core ideas and providing philosophical background on concepts, mental representation, similarity, and reasoning.

2.1. Core Ideas

Example. Gärdenfors's (1990, 2000, 2014) canonical example is the color space. It comprises three *quality dimensions* which represent color qualities objects may possess: hue, saturation, and brightness. Hue is a circular dimension capturing opponent colors such as red and green. Points in the space represent possible objects. The distance between them represents their (dis-)similarity. The *concept*, or *property* GREEN is identified with a region in the space.[1] To represent something as green means to map it into one's GREEN region.

Constructing CS. Building a conceptual space requires basic mathematical ingredients: axiomatic geometry ("*betweenness* and *equidistance* defined over a space of *points*" Gärdenfors 2000, p. 15), and a metric to measure distance (*Euclidian* or other, depending on the use case). Similarity is then defined as an exponentially decaying function of distance. Researchers select the quality dimensions. Here is a typical recipe: (i) collect pairwise similarity judgments (e.g., for color patches); (ii) locate the items in an *n*-dimensional space so that distances approximate the judged similarities by minimizing a "stress" loss (via multidimensional scaling, MDS); and (iii) choose the dimensionality with the lowest stress. The resulting axes are initially uninterpreted. In the case of the psychological color space, they can be interpreted as hue, saturation, and brightness, forming an inseparable package or *domain*.

Ontological Status. Gärdenfors (2000) views CS as "*theoretical entities* that can be used to explain and predict […]. [He eschews] philosophical discussions of how 'real' conceptual spaces are. The important thing is that we can *do* things with them" (p. 31). The bottom line is that CS are, first and foremost, versatile representational tools. Beyond describing cognition, they have been used as idealized *scientific* spaces which can represent theories in physics – and their change (Gärdenfors & Zenker 2010, 2013). They also serve as philosopher's tools for representing concepts such as identity and knowledge (Decock & Douven 2015), or even for analyzing currents of French feminism (Bendifallah 2023).

Importantly, however, it is also common practice to *ascribe* CS, regions, etc. to cognitive agents. Thus, it looks like CS are not only tools, but psychological postulates. In some sense, that is true. But I think we must be precise and say that we ascribe the *contents* which are represented by CS. Analogously, we would not want to ascribe a geographical map to the landscape it represents. But we would ascribe a river, for instance, of the same shape as the

[1] Gärdenfors distinguishes properties and concepts, but the difference is not important for the present discussion. I will simply speak of concepts, moving forward.

line which represents it on the map. The idea is that similarity is a real organizing principle of the mind, just like distance is a real organizing principle of the landscape – and formal geometry makes it precise. Thus, when we say that a system *has* a CS, or *has* a GREEN region by which it categorizes green objects, that is a shorthand for conceptual knowledge of a certain structure.

With that, let me address the frequently asked question whether one has a single, very large conceptual space, or many small spaces: Both is true. What I have is conceptual knowledge which is organized by similarity relations. Many different CS can justifiably be ascribed to me – as a shorthand – because my conceptual knowledge can be carved up in many different ways, just like a landscape can be represented by many different maps.

Cognitive economy and naturalness. An important line of thought is that the partitioning of the space into color concepts arises from principles of cognitive economy (Douven & Gärdenfors 2019). Because cognitive systems are limited (memory capacity, etc.), they continually optimize across competing demands. For example, they seek to extract as much information from the environment as possible (*informativeness*), but have to simplify due to their limitations (*parsimony*). The result of this balancing act is that the color concepts are *convex*: for any two points in a region, every point between them is also in that region. Thus, convexity is a geometric mark of cohesion. Gärdenfors treats convexity as necessary for naturalness. The optimality variables aim at a jointly sufficient account. That is, concepts are natural iff they are optimal (convex, easy to process, and so forth).

Mind-world relation. Words are thought not to refer directly to things in the external world, but to internal conceptual structures (*cognitive semantics*).[2] In philosophy of language, the notion that meanings are internal descriptions is known as *semantic internalism* (see Putnam 1975). It can lead to an undesirable *Kuhnian* relativism, meaning that knowledge, truth, etc. become relative to *incommensurable* theoretical paradigms (see Schrenk 2016, p. 238–239). Gärdenfors aims to avoid relativism by highlighting the social dimension of meaning and the way CS form in interaction with the word. Nevertheless, he opposes strong realism of any kind. His position may be characterized as *evolutionary pragmatism*, roughly: knowledge, truth, etc. are not arbitrary, but relative to our practical interests such as surviving and navigating a world that resists our actions. Papers (A) and (B) scrutinize this position.

[2] Gärdenfors (2014) elaborates the representational differences between various word classes, but this is orthogonal to the topics of this dissertation.

Individual differences in semantic knowledge. The framework accommodates individual differences. First, one can readily construct CS for different individuals or groups and compare them – again, they are versatile representational tools. Second, such differences have theoretical significance. Warglien & Gärdenfors (2013) characterize meaning construction as a "meeting of minds": an interactive, social process in which individuals search for fixed points – states "of mutual compatibility among individual strategies" (p. 2169). Put differently, groups with divergent conceptual structures attempt to establish common ground to communicate, coordinate, and navigate the world.

Empirical Corroboration. Empirically informed philosophers must work at the right level of abstraction. Their task is not to adjudicate the statistical details of individual studies, but to engage with the conceptual frameworks that guide both theories and experiments. Accordingly, this dissertation does not review the empirical CS literature; it proceeds in an agnostic – or better, conditional – mode: *assuming* that CS are empirically adequate, what follows for, say, inductive reasoning? This stance does not render empirical support irrelevant. On the contrary, empirical corroboration makes engagement with a framework worthwhile, and CS enjoy some notable support.[3] How much detail from particular models or studies is needed depends on the philosophical question at hand. The topics pursued here require relatively little detail, though papers (C) and (D) delve into cognitive modeling.

Beyond Gärdenfors. The CS framework has a rich history. Influences from theories of concepts, similarity, and reasoning are discussed in the next section. Here I highlight Rudolf Carnap, Roger Shepard, and the ubiquity of similarity spaces. Carnap is relevant because his broader project relies on similarity and, importantly, because he introduced *attribute spaces* in later work (1971, 1980), which anticipate CS. Within his project of developing an inductive logic, attribute spaces were used to specify relations among purely syntactic predicates (cf. Sznajder 2021, p. 468). Thus, partitioning into predicate-regions, a distance measure, and a version of cognitive semantics were already present. However, Carnap "never leaves the ideal that all knowledge should be expressed symbolically" (Gärdenfors 2011, p. 10), while Gärdenfors locates conceptual knowledge beneath language. Shepard (1987) is a critical influence from cognitive psychology. He models similarity as an exponentially decaying

[3] For a selection, I refer the interested reader to work on Shepard's law (Shepard 1987), RSA (Kriegskorte & Kievit 2013), optimal partitions of color space (Douven & Gärdenfors 2019, Regier et al. 2007), spatial codes (Bellmund et al. 2018, Constantinescu et al. 2016), graded membership (Douven 2016), category-based induction (Osta-Vélez & Gärdenfors 2020), robotics (Cubek et al. 2015), and convexity in deep networks (Tětková et al. 2025).

function of distance, which Gärdenfors adopts directly (2000, p. 20). He pioneered MDS for color spaces (p. 22) and "provides an evolutionary argument that supports the convexity criterion" (p. 70), among other contributions. Shepards work is one reason why similarity spaces are ubiquitous in modern cognitive science. Delineating CS from all these accounts is unfeasible, and not a goal of this thesis. Vicariously, however, I would like to mention *quality space theory* (QST; Clark 1992, Rosenthal 2010) – because it's discussed in philosophy. It "holds that the qualities presented by experiences in a given modality can be sorted in terms of their relative similarity" (Weger 2024, p. 12). The similarities with CS are striking. The main differences are, first, that QST focuses on the phenomenal character of experience, which is orthogonal to the CS account because CS explain and predict behavior, regardless of whether the agent has subjective experience, and second, that QST is not a theory of concepts. Gärdenfors (2000) references Clark, but does not engage QST substantively. A systematic comparison of CS with QST and other similarity space accounts remains a project for future research.

2.2. Concepts and Mental Representations

As noted above, concepts are commonly conceived as the building blocks of thought (Margolis & Laurence 2021). Philosophers have variously treated them as abstract objects,[4] sets of possible objects (e.g., Quine 1960), and more. Most contemporary philosophers of mind, however, view them as a special kind of *mental representation* (MR), posited and ascribed to explain higher cognitive capacities such as categorization or inductive reasoning. This stance aligns most naturally with the CS framework and is adopted here. Note, however, that CS is primarily a theory of the structure of conceptual content and is, in principle, compatible with multiple ontologies of concepts.[5]

The structure of conceptual content is a hotly debated issue since the classical Aristotelian view that concepts are lists of necessary and jointly sufficient conditions came under pressure. Fodor (1975) famously denied that concepts have internal structure, treating them as atomic symbols in a purely syntactic language of thought, a view that pushes toward "mad dog nativism" (Cowie 1998). By contrast, work in cognitive psychology has made it

[4] Frege, for instance, took thought contents to exist in a "third realm" beyond the physical and the mental (cf. Burge 1992).

[5] Gärdenfors 2000b only says that „meanings of linguistic expressions [i.e., concepts] are *mental entities*" (p. 3), which leaves some room for interpretation regarding the exact ontology.

common to treat concepts as structured. *Prototype* theory, for example, holds that concepts are organized around the best or most typical category members, a view motivated by well-established prototype effects – for instance, prototypicality reliably influences categorization latencies (Hampton 2006). Moreover, many natural categories exhibit prototypical structure (e.g., approximately normal distributions of biological features within populations), making prototype-based concepts unsurprising from an adaptive standpoint (Schurz 2012). Given a heterogeneous empirical landscape, some endorse pluralism (Weiskopf 2009) or even eliminativism about a unified concept of "concept" (Machery 2009; see Taylor & Vosgerau 2021 for discussion). A virtue of CS is that it accommodates at least prototypes and exemplars: both can be represented by points, with prototypes often – though not necessarily – identified as centers of gravity of regions. Given a set of prototype points, a convex partitioning can be generated via *Voronoi tessellation*, assigning every location to its nearest prototype.

Mental representation (MR) can be "broadly construed as a mental object with semantic properties" (Pitt 2020, p. 1), "being a vehicle and having a content" (Newen & Vosgerau 2020, p. 179) – which implies the possibility of misrepresentation. Contemporary discussion is heavily shaped by the idea that *intentionality*, or aboutness, is the hallmark of the mental (see Bartok 2005 for a discussion of Brentano's intentionality thesis). Many theories of MR aim at *naturalizing* intentionality, i.e., avoiding mind-body dualism, by connecting it with the semantic properties of representations.

> Some historical discussions of the representational properties of mind (e.g., [Aristotle, Locke, Hume]) seem to assume that nonconceptual representations – percepts ("impressions"), images ("ideas") and the like – are the only (or at least the main) kinds of mental representations, and that the mind represents the world in virtue of being in states that *resemble* things in it. (Ibid. sec 3).

Most contemporary accounts assume both conceptual and non-conceptual representations, while demarcation remains contested. One fruitful approach ties the distinction to the flexibility of the behavior to be explained (Newen & Vosgerau 2020). Some behaviors do not require positing representations (e.g., frogs reflexively snapping at flies and at black moving dots alike), whereas others do (e.g., ants start a special homing behavior if triangle completion fails). Vosgerau (2007) proposes ascribing conceptual representations when a system can "discriminate one property in different objects and more than one property in one object", which requires a "difference between the representation of objects and properties" (p. 350). CS meets this constraint by representing objects as points, and properties as regions or quality dimensions.

Gärdenfors (2000, sec. 2.3) relegates non-conceptual content to the *sub-symbolic* realm. Drawing an analogy: just as Locke and Hume emphasized associations among ideas, modern connectionist systems (e.g., deep neural networks) extract patterns from training data; their semantic content is distributed and high-dimensional. On the CS picture, spaces can be construed as arising via dimensionality reduction from sub-symbolic sensory information. At the same time, symbolic processing à la Fodor is not excluded: symbols can be mapped onto CS. In this way, CS *mediate* between sub-symbolic and symbolic levels, making Gärdenfors a pluralist about MR, though not about the format of conceptual content, which is geometric.

2.3. Similarity

Similarity is an extremely versatile concept with many meanings and use cases. Outside the CS debate, it is often conceived as a three-place relation of the form '*x* resembles *y* in respect *Q*', and with various axiomatic properties (symmetry, reflexivity, etc.; see Enflo 2020 for a detailed list and discussion). The geometric account defines it – and with it dissimilarity – as a two-place function of distance; the respect *Q* enters indirectly via the choice of quality dimensions.

A familiar philosophical use case occurs in metaphysics. Some positions take similarity as primitive (e.g., resemblance nominalism), while others analyze similarity in terms of more fundamental entities (e.g., *perfectly natural properties*, see Lewis 1983). Although these debates are not irrelevant (Lewis is discussed in papers A and B), Gärdenfors's treatment of similarity is explicitly non-, if not anti-, metaphysical. It is natural, however, to read CS as supporting *structural resemblance* – the idea that conceptual regions resemble what they represent – suggesting a correspondence between mind and world that might sit uneasily with anti-metaphysical scruples. As mentioned above, the resemblance idea traces back to Aristotle and Hume; contemporary authors often speak of second-order isomorphism (e.g., Edelman 1998). Gärdenfors (2000) instead speaks of intrinsic representations: "the representing relation has the same inherent constraints as its represented relation" (p. 44), a formulation that remains neutral about the metaphysics of the represented relation. The contrast class is symbolic representation, where meanings are arbitrarily and externally fixed. Metaphorically put, CS are less about the precise relation between the mirror and the world than about the structure within the mirror. A residual tension remains, and it is a formative theme of papers (A) and (B).

An important landmark in the philosophical debate on similarity is Carnap's (1967) *Aufbau*. As a logical empiricist, Carnap was likewise skeptical of metaphysics. Thus, he sought to reconstruct scientific properties from elementary experiences and a "recollection of similarity" (p. 127) relation defined over them.[6] The reconstruction method – *quasi-analysis* – need not be rehearsed here. Goodman (1966, 1972) famously criticized Carnap and the very idea that similarity could constitute properties, convincing many that similarity is philosophically idle. The details of Goodman's arguments are not required here; what matters is that they target a conception of similarity different from the geometric account used in CS. For example, Goodman's claim that "any two objects share at least one property, so that similarity is a universal relation – everything is similar to everything else – and similarity claims are thus utterly uninformative" (Decock & Douven 2011, p. 68) is based on the suppositions that properties are sets of objects, and that similarity should be context-independent. CS avoids this result by treating properties as regions – sets of points representing possible objects – and by encoding contextual factors in the quality dimensions. Subsequent progress has yielded a partial rehabilitation of similarity in philosophy, and even of quasi-analysis (see Decock & Douven 2011).

In cognitive science, similarity is a frequent guest in psychological explanations of cognitive capacities, both as a useful modeling tool, and as a foundational organizing principle of the mind (Goldstone & Son 2012). Geometric similarity, popularized by Shepard and later Gärdenfors, is one influential conception among others. Its main competitor is Tversky's (1977) *set-theoretic* account, which emphasizes shared versus distinctive features and the role of contrast classes. This dissertation does not engage competing conceptions in detail, but one difference is salient: the set-theoretic approach operates on discrete features, whereas CS quality dimensions may be discrete but are often continuous. A general advantage of continuity is the ability to capture fine-grained contents, especially in perception.

2.4. Reasoning

A full treatment of reasoning is impossible here, so I highlight only ideas directly relevant to CS. First, reasoning is closely tied to *rationality*, which implies that there are better and worse (rational vs. irrational) ways to reason, making the topic one of *normative epistemology*. What counts as rational is evaluated against benchmarks (Schurz & Hertwig 2019). While such

[6] On a similar note, Quine (1969) argued that natural kinds are a matter of similarity, see papers (A) and (B).

benchmarks were often conceived as universal, some contemporary views – e.g., ecological rationality (Gigerenzer & Todd 1999) – treat them as local and context-sensitive, a move critics warn risks an is-ought slide (Elqayam & Evans 2013). Instrumental considerations about which means best achieve a given benchmark "involve a normative dimension insofar as [they shift] the normative weight from the end to the means" (Schurz & Hertwig 2019, p. 11). In this sense, the CS framework is normative insofar as it engages with optimality relative to goals such as informativeness and parsimony. That said, claims about what is optimal given a goal can also be framed descriptively:

> [...] normative epistemology is a branch of engineering. It is the technology of truth-seeking [...] There is no question here of ultimate value […]; it is a matter of efficacy for an ulterior end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed. (Quine 1986: 664–665).

For that matter, Strößner (2023) speaks of "hybrid frameworks" in a complex "normative-descriptive interplay". This dissertation reserves the predicate 'normative' for accounts which concern the justification of the benchmarks of reasoning (e.g., Schurz & Hertwig's justification by *cognitive success*). How CS might justify aspects of reasoning is discussed in paper (A).

At its core, reasoning involves *inference*. *Inductive generalization* is *uncertain* inference from observations to general law hypotheses. "The induction principle asserts that regular connections between events that have been observed in the past will also hold in the future; but Hume argued that this inference can be justified neither by logic nor by experience" (Schurz 2014, p. 5). *Abductive inference*, or inference to the best explanation (IBE), is, "[r]oughly expressed, […] an inference from an observed effect to a likely cause, e.g. from a curved trail in the sand to the recent passage of a sand viper" (ibid., p. 55). Although ubiquitous in the sciences, abductive inference is provisional and contested, and what counts as the best explanation will of course heavily depend on what an explanation is, and what the criteria for good explanations are (see section 4). *Deductive inferences* "transfer the truth from the premises to the conclusion with certainty" (ibid., p. 50); they are the object of logic in the narrow sense. Due to its rigor, deduction was long viewed as the paradigm of rational reasoning and the normative benchmark of choice – in philosophy and cognitive psychology (cf. Wason 1969). Its chief historical rival, probabilistic inference, is introduced below; additional forms – such as analogical reasoning – are treated in the papers.

3. Papers

> Via successful and less successful interactions with the world, the conceptual structure of an individual will adapt to the structure of reality. It must be emphasized, however, that this does not entail that the conceptual structure represents the world. Gärdenfors 2000, p. 156.

This quotation embodies the evolutionary pragmatism which shapes Gärdenfors's metaphysical and epistemological outlook – especially regarding naturalness – and which is scrutinized in papers (A) and (B). It also foregrounds interaction as central to his philosophy, motivating papers (C) and (D). Examining dynamic and interactive models ultimately raises the question of what we expect from such models; the answer turns on explanation, the focus of paper (D). Section 4 returns to this theme. I now turn to the individual articles.

3.1. (A) Goodman's Riddle

Gärdenfors (1990) presents the CS account as a comprehensive solution to Goodman's (1983) *New Riddle of Induction*, which famously undermines the justifiability of our reasoning practices. Goodman's own solution is *instrumentalist*. The paper argues that Gärdenfors's solution strategy collapses into Goodman's because his realism about naturalness is too weak. Thus, progress is made only in the description, not the justification of induction. The point requires some background on the realism debate.

3.1.1. The Realism Debate

Realism is a broad term. Unqualified – *global* realism – it typically holds that an external, mind-independent world exists and causes our experiences. *Local* realisms concern specific domains, e.g., realism about universals (multiply instantiable entities) or Platonism (non-spatiotemporal entities; Rodriguez-Pereyra 2019). *Anti-realism* denies such commitments; for example, *nominalism* holds that only concrete particulars exist, though most nominalists remain robust realists about those particulars.

Scientific anti-realists hold that "science is possible if it is merely based on human intersubjective experiences, without the presupposition of a subject-independent reality that transcends these experiences." (Schurz 2014, p. 59–60). For example, conventionalists about *natural kinds* (e.g., gold, *Amanita muscaria*) regard scientific classification as a matter of convention (Bird & Tobin 2024). Because this seems to make discovery a rather capricious affair, many realists worry that it invites problematic relativism. *Instrumentalists* reject realism yet hold that theories and kinds may be empirically adequate ("therefore also in the future

the most empirically successful" Schurz 2014, p. 57), i.e., science is not capricious. For present purposes, *pragmatists* can be read as instrumentalists who accept that science is relative to practical interests. These latter positions may be called weakly realistic. By contrast, strong scientific realists infer from the empirical success of theories that they are true or truthlike – a form of abductive inference: without realism, the success of science would be miraculous.

Gärdenfors argues against strong realism about natural kinds, essences, causes, and other realist favorites, instead offering cognitivist analyses that make such notions relative to our conceptual structures. To avert problematic relativism, he emphasizes that our conceptual structures develop through interaction with the external world – the point of the quotation above. In a footnote to that passage, he identifies as Neo-Kantian: the world in itself is epistemically inaccessible. He adopts a pragmatist stance that foregrounds our evolutionary interests ("adapt[ation] to the structure of reality"). The footnote suggests that the passage should not be read as a rejection of mental representation in cognitive science, but as a rejection of a strong, pre-Kantian mirroring realism.

3.1.2. Results

The paper analyzes Gärdenfors's strategy as an attempt to ground the instrumental success of certain predicates (e.g., GREEN rather than Goodman's infamous GRUE) in their naturalness, thereby justifying their use in induction. Strong realists typically pursue such a strategy by appealing to the joints of nature, mind-independent and objective. Gärdenfors, by contrast, locates naturalness at the joints of thought. Combined with evolutionary pragmatism – "creatures inveterately wrong in their inductions have a pathetic, but praiseworthy tendency to die before reproducing their kind" (Quine 1969, p. 126) – this approach approximates Goodman's view that we simply retain predicates that *work* (help us survive and navigate). Moreover, appeals to evolution are problematic: poor reasoning can be adaptive, yielding a skeptical conclusion. The paper concludes that the abductive no-miracles argument leads out of the skepticism, and that Gärdenfors must endorse a more thorough realism for his solution strategy to work.

3.2. Cognitive Sparseness

All this merits an even closer look at the cognitive and ontological dimensions of naturalness. The philosophical debate on naturalness is split into three camps. (i) Some treat natural kinds

as essences, motivated by semantic externalism's demand for modally stable referents of kind terms. (ii) Others posit natural properties and kinds as primitives to underwrite metaphysical accounts of laws and causation. (iii) A third group invokes natural kinds to ground induction and scientific practice. CS aligns most closely with (iii), yet evolutionary pragmatism undercuts the required objectivity – strikingly, CS would count *fictional kinds* (orcs, gods, the creatures of legends and myths) as natural, a highly revisionary implication. Two distinct problems arise: the *problem of sparseness*, concerning how many predicates are admissible in reasoning; and the *problem of objectivity*, concerning which objective entities those predicates refer to. CS addresses the first but not the second; hence, cognitive naturalness is better labeled *cognitive sparseness*. Unlike the first paper, the second identifies the root difficulty not in weak realism per se, but in the evolutionary component of the pragmatism. Evolution is too broad and unreliable to secure a stable connection to the world – and thus to objectivity. A *methodological* form of objectivity can, however, be achieved by appealing to the reliability of scientific method (Sankey 2024). The paper advocates *ecological empiricism*, a pragmatic view on which naturalness arises from interaction between agents and world (Vosgerau 2024). When interaction takes the form of scientific measurement and theorizing, it yields methodologically objective kinds.

3.3. Predictive Models

Although interaction is at the heart of Gärdenfors's epistemology, as evidenced by the quote, the CS account is static in the sense that it maps the structure of semantic knowledge at a given time. And although he discusses concept acquisition and learning, a clear picture of CS dynamics is still missing. Given the shortcomings of the evolutionary conception of interaction, it is natural to seek a more precise alternative. A thriving paradigm placing interaction center stage is Predictive Processing (PP), which will be outlined below. Ecological empiricism can be framed as a version of PP that does not commit to PP's full formalism but highlights *affordances* (roughly, opportunities for action). Consequently, paper (C) investigates the relationship between CS and PP. Some background on one of PP's central notions – probability – is required.

3.3.1. Probability

No detailed introduction to the formalism or history of probability theory is required here. Instead, we need to know what the papers are talking about, what not, and how probability is significant for the issues at hand. What they are generally *not* talking about is objective probability, or relative frequency, that is, the probabilities of worldly events. Instead, we are interested in how probability is represented – the cognitive dimension of probability. Thus, the papers speak of psychological probability, or *subjective degrees-of-belief* (cf. Schurz 2015). While Gärdenfors himself is not a paragon of probability, this approach comports with his Neo-Kantian views on the epistemic accessibility of (relative frequencies in) the world in itself. Our access is indirect: We estimate relative frequencies by updating our subjective degrees-of-belief as we interact with the world and make observations.[7] This can be expressed in terms of *posterior* conditional probabilities: how likely a hypothesis is given the evidence. *Bayes' theorem*, derivable from the probability axioms, specifies how such posteriors ought to be updated. Papers (C) and (D) therefore explore connections between CS and *Bayesianism*.

There are several well-known (though contested) *a priori* arguments for Bayesianism. For example, agents whose credence functions violate the probability axioms are susceptible to *Dutch book* arrangements – bets that guarantee a loss. For this and other reasons, many twentieth-century philosophers and psychologists treated probabilistic reasoning as a normative benchmark – a successor to deductive reasoning. Yet cognitive psychology shows that real agents often violate probabilistic norms, motivating ecological rationality and related approaches. Despite this, Bayesian methods remain influential in epistemology and across the sciences – and, crucially, have been powerful tools in cognitive modeling:

> [I]t has been applied to the problem of perception and categorisation and the problem of sensory coding and processing. The idea that perception should be understood as being inferential is often traced back to Helmholtz. Already Helmholtz used probabilities to model the nature of the unconscious inferential processes underlying perception and categorization. […]. The more general Bayesian brain hypothesis, i.e. the hypothesis that in general (not just in perception) the brain codes and processes as if it were employing Bayesian methods or models, is discussed by philosophers of cognition and neuroscientists […]. Brössel 2017, p. 733–734.

[7] This is not meant to imply that science must do without objective probability, but "objective" would have to be understood as the instrumentalistic, methodological objectivity advocated in paper (B). Schurz (2015) develops a dualist view which aims to reconcile the subjective with the objective conception. Given that he justifies norms of reasoning by appealing to their cognitive success (Schurz & Hertwig 2019), it is *prima facie* compatible with the account presented here – but this issue is beyond the scope of this thesis.

The Bayesian brain hypothesis is a core tenet of PP. The latter is sometimes presented as a "grand unified theory" (Clark 2013, p. 200) of cognition built on Bayesian principles. With this, the discussion shifts from issues in metaphysics and philosophy of science to the philosophy of cognition.

3.3.2. Results

Alongside the Bayesian brain hypothesis, PP emphasizes *prediction error minimization*, *hierarchical generative models*, and *active inference*. Conceptual spaces that instantiate these tenets can function as *predictive models*. There is no *prima facie* reason CS cannot do so, but compatibility constraints on the goals, ontology, and methodology of the theories must be met. For example, the version of PP to be combined with CS must not be radically *enactivist*, which would amount to it rejecting mental representations. One potential route to compatibility would be to reduce psychological similarity to subjective probability, or conversely. The paper argues, following Poth (2019), that such a reduction fails because similarity has semantic value for interpreting Bayesian models (see section 3.4).

Recent models demonstrate *partial methodological compatibility* by combining CS with subjective degrees-of-belief functions. Douven et al. (2025), for instance, start with a flat prior over a space, and model the belief update with a weighted Gaussian mixture, meaning that new observations are associated with a Gaussian distribution that is weighted and added to the prior distribution.[8] Other accounts apply Bayes' theorem directly. Together, these models show how CS can fit within the broader Bayesian brain picture. In this way, paper (C) articulates non-evolutionary interaction between CS and world: cognitive systems use geometric conceptual knowledge to generate predictions, actively test them, and update rational degrees-of-belief functions defined over that conceptual knowledge.

A major open question concerns structural updating: how should the *geometry of the space itself* change when new conceptual knowledge is acquired? Moreover, a hallmark of PP not yet captured by current hybrids is the hierarchical generative organization of the brain's models. Developing a hierarchical, generative CS with mechanisms for structure learning is a promising direction for future work.

[8] The fact that Douven et al. use a flat prior in accordance with the *principle of indifference* makes them *objective Bayesians*, but that doesn't mean that the represented probabilities are relative frequencies (see Schurz 2015, sec. 9.3).

3.4. Semantically Deficient Models of Cognition

The final paper takes a modest step toward such a model – though only in its last section and under the heading of *semantic opaqueness*. Its main theme is Poth's (2019, 2023) aforementioned critique of Bayesian cognitive models, exemplified by Tenenbaum & Griffiths' (2001) model of perceptual generalization. She argues that geometric similarity is valuable for semantically interpreting the so-called "size principle" – a rational preference for smaller hypotheses – and the beliefs in the Bayesian model. Thereby, she touches on the question of what CS contribute in an era of increasingly complex, predictively powerful cognitive models – specifically, what advantages CS offer cognitive scientists. The focus shifts from the philosophy of cognition to the philosophy of cognitive science, clarifying the backdrop of Poth's critique, her proposed remedy, and implications for the hybrid models in paper (C).

The main results are: 1. Poth's critique is related to, but different from the hotly debated notion of AI opacity. To avoid confusion, the paper suggests "Semantically Deficient Models of Cognition", or "SEDMOC", as improved terminology. 2. Her critique rests on the idea that explanatory cognitive models should indicate how cognitive content is efficacious in producing the target behavior. Semantically opaque prediction engines fail this test. 3. The demand for model interpretability coheres with a pragmatic theory of explanation. 4. CS prevent semantic opaqueness by providing intrinsically meaningful representations, thereby avoiding the *just-so trap* and the *symbol grounding problem* – the two facets of Poth's critique. 5. The hybrid models from paper (C) avoid these problems, but their cognitive adequacy improves with a clearer picture of CS dynamics. 6. A PCA-based abstraction mechanism is introduced as a cognitively economic heuristic for shaping higher layer spaces in hierarchical CS. The twist is that the model doesn't need to "remember" individual observations, but can use the subjective degrees-of-belief function to generate a sample which serves as the input for PCA – a statistical method for capturing covariation. 7. The proposed abstraction mechanism can introduce semantically opaque dimensions. This is not detrimental, however, because even if linguistic labels cannot attach to the dimensions, they can be traced to interpretable dimensions in lower-layers.

The take home message with regard to the philosophy of cognitive science is that CS have a higher explanatory potential than semantically opaque models of the same predictive

power because they commit to intrinsically meaningful psychological entities (regions, dimensions, and so forth).

4. Developments and Results

Paper (A) centers on naturalness and Goodman's riddle. It operates at the intersection of cognitive science, philosophy of mind, metaphysics of science (natural kinds, realism), philosophy of language (cognitive semantics; semantic externalism vs. internalism), philosophy of science (the riddle of induction), and epistemology (rationality). Broadly the same constellation applies to the other papers, with shifting priorities. With PP, the philosophy of cognition takes center stage in paper (C), while paper (D) addresses the philosophy of cognitive science by examining the explanatory value of cognitive modeling. It is fair to say that the thematic development of the thesis testifies to the versatility and philosophical fruitfulness of the CS framework – an unsurprising, but important, general result.

The strong realism defended in paper (A) shifts in paper (B) toward a more pragmatic stance. I initially adopted the former because Gärdenfors's position appeared unstable: at times he distances himself from realism, yet he also endorses strongly realist authors such as Richard Boyd and Hilary Kornblith (Gärdenfors & Stephens 2017). This ambiguity suggested to me that strong realism was doing the heavy lifting in avoiding relativism. With the discussion of objectivity in paper (B), however, a more nuanced understanding of how instrumentalists and pragmatists avoid problematic relativism emerged. Roughly, our access to the world is mediated by our concepts, and abductive inferences to real kinds, etc. are contentious (see below). Yet by making our interaction with the world as systematic and objective as possible – i.e., by refining scientific methods – we may not secure truth with a capital *T*, but we can track the world's structure with sufficient reliability. The difficulty is that we cannot know whether we are close to the truth or "forever stuck in a local optimum" (Douven 2024, p. 854).

Skepticism toward abductive (or "no-miracles") arguments for realism is reinforced by two considerations not developed in the papers. First, a clearer view of the problems of abduction, especially the status of explanation. Although paper (D) does not survey theories of explanation in depth, the literature offers no consensus on what explanations are, let alone what would make one the "best." If the line pursued there is correct, irreducible contextual factors are always in play, undermining the idea of a best explanation in any absolute sense.

Second, historical work on temperature and its measurement.[9] Chang (2004) shows convincingly that establishing fixed points (e.g., the boiling point of water) *independently* is impossible, because the reliability of any measurement is always relative to other measurements and to the surrounding theoretical and conceptual frameworks. As Chang puts it, "[w]e would like to put some nails in the wall to hang things from, but there is actually no wall there yet" (p. 40). Accordingly, the history of the boiling point unfolds as an open-ended, iterative process in which scientists identify interfering factors (contamination, surface adhesion, etc.) to secure increasingly stable correlations. Such work produces robust, reliable knowledge, but not truths about the world "in itself." In this sense, Chang offers a practice-oriented illustration of a broadly Kantian, anti-metaphysical stance – except that the "categories" of our thinking (here interpreted as CS) are themselves malleable. Chang develops a form of coherentism from these observations, which I set aside here. The position of the thesis can be summarized as follows.

Naturalness. Methodologically objective natural kinds arise from the interaction between cognition and the world; more specifically from systematic interaction guided by continuously refined scientific methods. Cognitive criteria of naturalness such as convexity in CS are only one part of the story, namely, the sparseness-part. The objectivity-part comes from appealing to the nature of the interaction, more specifically the success of science. Gärdenfors's appeal to evolution cannot establish objectivity because objectively poor reasoning can be adaptive. Metaphysical theories of naturalness address the sparseness-part, too, but their criteria are not cognitively motivated. Thus, granted their empirical adequacy, Gärdenfors's cognitive sparseness criteria are a major advancement for the naturalness debate overall, but not for the metaphysical problems discussed in that debate – as long as they are tied to *evolutionary* pragmatism.

Interaction. The results on naturalness highlight the general epistemological significance of interaction, but a clear picture of how CS change dynamically through interaction is still missing. PP can provide such a picture if CS can function as predictive models. Recent work suggests they can: agents may update a degree-of-belief function that ranges over their geometric, conceptual representations and use it to generate predictions for interactive error minimization. The next step is to create hierarchical generative CS models with an update

[9] I am grateful to Gottfried Vosgerau for bringing up this example in a conversation.

mechanism for the structure of the underlying space. The PCA based abstraction mechanism introduced in paper (D) serves as a starting point.

Explanation. Abductive arguments for scientific realism – endorsed in paper (A) – hinge on the notion of a best explanation, which in turn depends on a theory of explanation. A full treatment lies beyond the scope of this thesis, but the provisional pragmatism I adopt and defend in paper (D) sits uneasily with no-miracles arguments. The principal result of paper (D) concerning explanation is that explanatory models in cognitive science should make explicit how cognitive contents affect the behaviors to be explained. Due to their semantic transparency, CS thereby offer explanatory value in a landscape increasingly populated by predictive – but semantically opaque – models of cognition.

References

Bartok, Philip J. 'Brentano's Intentionality Thesis: Beyond the Analytic and Phenomenological Readings'. *Journal of the History of Philosophy* 43, no. 4 (2005): 437–60. https://doi.org/10.1353/hph.2005.0153.

Bellmund, Jacob L. S., Peter Gärdenfors, Edvard I. Moser, and Christian F. Doeller. 'Navigating Cognition: Spatial Codes for Human Thinking'. *Science* 362, no. 6415 (2018): eaat6766. https://doi.org/10.1126/science.aat6766.

Bendifallah, Lina, Julie Abbou, Igor Douven, and Heather Burnett. 'Conceptual Spaces for Conceptual Engineering? Feminism as a Case Study'. *Review of Philosophy and Psychology* 16, no. 1 (2025): 199–229. https://doi.org/10.1007/s13164-023-00708-7.

Bird, Alexander, and Emma Tobin. 'Natural Kinds'. In *The Stanford Encyclopedia of Philosophy*, Spring 2024, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University, 2024. https://plato.stanford.edu/archives/spr2024/entries/natural-kinds/.

Brössel, Peter. 'Rational Relations Between Perception and Belief: The Case of Color'. *Review of Philosophy and Psychology* 8, no. 4 (2017): 721–41. https://doi.org/10.1007/s13164-017-0359-y.

Brössel, Peter, Anna-Maria A. Eder, and Franz Huber. 'Evidential Support and Instrumental Rationality'. *Philosophy and Phenomenological Research* 87, no. 2 (2013): 279–300. https://doi.org/10.1111/j.1933-1592.2011.00543.x.

Burge, Tyler. 'Frege on Knowing the Third Realm'. *Mind* 101, no. 404 (1992): 633–50.

Carnap, Rudolf. 'A Basic System of Inductive Logic, Part I'. In *Studies in Inductive Logic and Probability*, edited by Richard C. Jeffrey. University of California Press, 1971.

———. 'A Basic System of Inductive Logic, Part II'. In *Studies in Inductive Logic and Probability*, edited by Richard C. Jeffrey. University of California Press, 1980.

———. *The Logical Structure of the World*. Edited by Rudolf Carnap. University of California Press, 1967.

Chakravartty, Anjan. 'On the Prospects of Naturalized Metaphysics'. In *Scientific Metaphysics*, edited by Don Ross, James Ladyman, and Harold Kincaid. Oxford University Press, 2013.

———. 'Scientific Realism'. In *The Stanford Encyclopedia of Philosophy*, Summer 2017, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2017. https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/.

Chang, Hasok. *Inventing Temperature: Measurement and Scientific Progress*. OUP Usa, 2004.

Clark, Andy. 'Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science'. *Behavioral and Brain Sciences* 36, no. 3 (2013): 181–204. https://doi.org/10.1017/s0140525x12000477.

Clark, Austen. *Sensory Qualities*. Clarendon Press, 1992.

Constantinescu, Alexandra O., Jill X. O'Reilly, and Timothy E. J. Behrens. 'Organizing Conceptual Knowledge in Humans with a Grid-like Code'. *Science (New York, N.Y.)* 352, no. 6292 (2016): 1464–68. https://doi.org/10.1126/science.aaf0941.

Cowie, Fiona. 'Mad Dog Nativism'. *The British Journal for the Philosophy of Science* 49, no. 2 (1998): 227–52. https://doi.org/10.1093/bjps/49.2.227.

Cubek, Richard, Wolfgang Ertel, and Günther Palm. *High-Level Learning from Demonstration with Conceptual Spaces and Subspace Clustering*. In *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015. 2015. https://doi.org/10.1109/ICRA.2015.7139548.

Decock, Lieven, and Igor Douven. *Conceptual Spaces as Philosophers' Tools*. 2015. https://doi.org/10.13140/2.1.1441.6966.

Douven, Igor. ‘Abduction’. In *The Stanford Encyclopedia of Philosophy*, Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/sum2021/entries/abduction/.

———. ‘Naturalness, Scientific Concepts, and the Substantivity of Social Metaphysics’. *Philosophia* 52, no. 4 (2024): 849–63. https://doi.org/10.1007/s11406-024-00776-8.

———. ‘Vagueness, Graded Membership, and Conceptual Spaces’. *Cognition* 151 (June 2016): 80–95. https://doi.org/10.1016/j.cognition.2016.03.007.

Douven, Igor, and Peter Gärdenfors. ‘What Are Natural Concepts? A Design Perspective’. *Mind and Language*, no. 3 (2019): 313–34. https://doi.org/10.1111/mila.12240.

Douven, Igor, Steven Verheyen, Shira Elqayam, Peter Gärdenfors, and Matías Osta-Vélez. ‘Analogical Reasoning: A Carnapian Approach’. *Synthese* 205, no. 5 (2025): 1–34. https://doi.org/10.1007/s11229-025-04991-y.

Edelman, Shimon. ‘Representation Is Representation of Similarities’. *Behavioral and Brain Sciences* 21, no. 4 (1998): 449–67. https://doi.org/10.1017/s0140525x98001253.

Elqayam, Shira, and Jonathan Evans. ‘Rationality in the New Paradigm: Strict versus Soft Bayesian Approaches’. *Thinking and Reasoning* 19 (July 2013): 453–70. https://doi.org/10.1080/13546783.2013.834268.

Enflo, Karin. ‘Measures of Similarity’. *Theoria* 86, no. 1 (2020): 73–99.

Fodor, J.A. *The Language of Thought*. Crowell, 1975.

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

———. *Does Semantics Need Reality?* 2000b. https://doi.org/10.1007/978-0-585-29605-0_23.

———. ‘Induction, Conceptual Spaces and AI’. *Philosophy of Science* 57, no. 1 (1990): 78–95.

———. *Inductive Reasoning: From Carnap to Cognitive Science*. 1 January 2011.

———. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press, 2014.

Gärdenfors, Peter, and Andreas Stephens. ‘Induction and Knowledge-What’. *European Journal for Philosophy of Science* 8, no. 3 (2017): 471–91. https://doi.org/10.1007/s13194-017-0196-y.

Gärdenfors, Peter, and Frank Zenker. ‘Theory Change as Dimensional Change: Conceptual Spaces Applied to the Dynamics of Empirical Theories’. *Synthese* 190, no. 6 (2013): 1039–58. https://doi.org/10.1007/s11229-011-0060-0.

———. *Using Conceptual Spaces to Model the Dynamics of Empirical Theories*. 2010. https://doi.org/10.1007/978-90-481-9609-8_6.

Goodman, Nelson. *Fact, Fiction, and Forecast*. 4th Edition. Harvard University Press, 1983.

———. ‘Seven Strictures on Similarity’. In *Problems and Projects*. Bobbs-Merrill, 1972.

———. *The Structure of Appearance*. Second. Bobbs-Merrill, 1966.

Kriegeskorte, Nikolaus, and Rogier A. Kievit. ‘Representational Geometry: Integrating Cognition, Computation, and the Brain’. *Trends in Cognitive Sciences* 17, no. 8 (2013): 401–12. https://doi.org/10.1016/j.tics.2013.06.007.

Lewis, David. ‘New Work for a Theory of Universals’. *Australasian Journal of Philosophy* 61, no. 4 (1983): 343–77. https://doi.org/10.1080/00048408312341131.

Machery, Edouard. *Doing without Concepts*. Oxford University Press, 2009. https://doi.org/10.1093/acprof:oso/9780195306880.001.0001.

Margolis, Eric, and Stephen Laurence. ‘Concepts’. In *The Stanford Encyclopedia of Philosophy*, Spring 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/spr2021/entries/concepts/.

Newen, Albert, and Gottfried Vosgerau. ‘Situated Mental Representations: Why We Need Mental Representations and How We Should Understand Them’. In *What Are Mental Representations?*, by Albert Newen and Gottfried Vosgerau. Oxford University Press, 2020. https://doi.org/10.1093/oso/9780190686673.003.0007.

Osta-Vélez, Matías, and Peter Gärdenfors. ‘Category-Based Induction in Conceptual Spaces’. *Journal of Mathematical Psychology* 96 (June 2020): 102357. https://doi.org/10.1016/j.jmp.2020.102357.

Poth, Nina. ‘Conceptual Spaces, Generalisation Probabilities and Perceptual Categorisation’. In *Conceptual Spaces: Elaborations and Applications*, edited by Peter Gärdenfors, Antti Hautamäki, Frank Zenker, and Mauri Kaipainen. Springer Verlag, 2019.

———. ‘Probabilistic Learning and Psychological Similarity’. *Entropy* 25, no. 10 (2023): 1407. https://doi.org/10.3390/e25101407.

Putnam, Hilary. ‘The Meaning of “Meaning”’. *Minnesota Studies in the Philosophy of Science* 7 (1975): 131–93.

Quine, W. V. ‘Reply to Morton White’. In *The Philosophy of W. V. Quine*, edited by L. Hahn and P. Schilpp. Open Court, 1986.

———. *Word and Object*. MIT Press, 1960.

Regier, Terry, Paul Kay, and Naveen Khetarpal. ‘Color Naming Reflects Optimal Partitions of Color Space’. *Proceedings of the National Academy of Sciences* 104, no. 4 (2007): 1436–41. https://doi.org/10.1073/pnas.0610341104.

Rodriguez-Pereyra, Gonzalo. ‘Nominalism in Metaphysics’. In *The Stanford Encyclopedia of Philosophy*, Summer 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2019. https://plato.stanford.edu/archives/sum2019/entries/nominalism-metaphysics/.

Rosenthal, David. ‘How to Think About Mental Qualities’. *Philosophical Issues* 20, no. 1 (2010): 368–93. https://doi.org/10.1111/j.1533-6077.2010.00190.x.

Sankey, Howard. ‘The Objectivity of Science’. *Journal of Philosophical Investigations at University of Tabriz* 17, no. 45 (2023): 1–10. https://doi.org/10.22034/jpiut.2023.17473.

Scholz, Sebastian. ‘Semantically Deficient Models of Cognition – A Bayesian Case Study, Conceptual Spaces, and AI Opacity’. *Review of Philosophy and Psychology*, ahead of print, 7 August 2026b. https://doi.org/10.1007/s13164-026-00834-y.

———. ‘Conceptual Spaces as Predictive Models’. *Synthese* 207, no. 6 (2026a): 237. https://doi.org/10.1007/s11229-026-05549-2.

———. ‘Conceptual Spaces: A Solution to Goodman’s New Riddle of Induction?’ *Philosophia*, ahead of print, 23 July 2024. https://doi.org/10.1007/s11406-024-00744-2.

Scholz, Sebastian, and Gottfried Vosgerau. ‘Conceptual Spaces: Naturalness or Cognitive Sparseness?’ *Synthese* 205, no. 3 (2025): 129. https://doi.org/10.1007/s11229-025-04941-8.

———. ‘The Cognitive and Ontological Dimensions of Naturalness – Editor’s Introduction’. *Philosophia* 52, no. 4 (2024): 845–48. https://doi.org/10.1007/s11406-024-00775-9.

Schrenk, Markus. *Metaphysics of Science: A Systematic and Historical Introduction*. Routledge, 2016. https://doi.org/10.4324/9781315639116.

Schurz, Gerhard. ‘Prototypes and Their Composition From an Evolutionary Point of View’. In *The Oxford Handbook of Compositionality*, edited by Markus Werning, Wolfram Hinzen, and Edouard Machery. Oxford University Press, 2012.

———. *Philosophy of Science: A Unified Approach*. Routledge, 2014.

———. *Wahrscheinlichkeit*. DE GRUYTER, 2015. https://doi.org/10.1515/9783110420364.

Schurz, Gerhard, and Ralph Hertwig. ‘Cognitive Success: A Consequentialist Account of Rationality in Cognition’. *Topics in Cognitive Science* 11, no. 1 (2019): 7–36. https://doi.org/10.1111/tops.12410.

Shepard, R. N. ‘Toward a Universal Law of Generalization for Psychological Science’. *Science (New York, N.Y.)* 237, no. 4820 (1987): 1317–23. https://doi.org/10.1126/science.3629243.

Strößner, Corina. ‘Hybrid Frameworks of Reasoning: Normative-Descriptive Interplay’. *Proceedings of the Annual Meeting of the Cognitive Science Society* 45, no. 45 (2023). https://escholarship.org/uc/item/58j925zr.

Taylor, Samuel D., and Gottfried Vosgerau. ‘The Explanatory Role of Concepts’. *Erkenntnis* 86, no. 5 (2021): 1045–70. https://doi.org/10.1007/s10670-019-00143-0.

Tenenbaum, Joshua B., and Thomas L. Griffiths. ‘Generalization, Similarity, and Bayesian Inference’. *Behavioral and Brain Sciences* 24, no. 4 (2001): 629–40. https://doi.org/10.1017/s0140525x01000061.

Tětková, Lenka, Thea Brüsch, Teresa Karen Scheidt, et al. ‘On Convex Decision Regions in Deep Network Representations’. *Nature Communications* 16, no. 1 (2025): 5419. https://doi.org/10.1038/s41467-025-60809-y.

Tversky, Amos. ‘Features of Similarity’. *Psychological Review* (US) 84, no. 4 (1977): 327–52. https://doi.org/10.1037/0033-295X.84.4.327.

Vosgerau, Gottfried. ‘Conceptuality in Spatial Representations’. *Philosophical Psychology* 20, no. 3 (2007): 349–65. https://doi.org/10.1080/09515080701197130.

———. ‘Ecological Empiricism’. *Philosophia*, ahead of print, 30 May 2024. https://doi.org/10.1007/s11406-024-00740-6.

———. ‘Varieties of Representation’. In *Knowledge and Representation*, edited by A. Newen, A. Bartels, and E. Jung. CSLI Publications, mentis, 2011.

Warglien, Massimo, and Peter Gärdenfors. ‘Semantics, Conceptual Spaces, and the Meeting of Minds’. *Synthese* 190, no. 12 (2013): 2165–93.

Wason, P. C. ‘Regression in Reasoning?’ *British Journal of Psychology* 60, no. 4 (1969): 471–80. https://doi.org/10.1111/j.2044-8295.1969.tb01221.x.

Weger, Daniel Mario. ‘Representationalism and Molyneux’s Question: An Intermodal Approach Based on Quality Space Theory’. *Philosophy and the Mind Sciences* 5 (December 2024). https://doi.org/10.33735/phimisci.2024.11601.

Weiskopf, Daniel Aaron. ‘The Plurality of Concepts’. *Synthese* 169, no. 1 (2009): 145–73. https://doi.org/10.1007/s11229-008-9340-8.

AI disclosure. ChatGPT assisted with idea clarification and language editing. I verified all claims and citations independently; any retained wording was edited and is supported by the listed references.

Part II – Publications

Paper A

Conceptual Spaces: A Solution to Goodman's New Riddle of Induction?

Bibliographic reference. Scholz, S. (2024). Conceptual Spaces: A Solution to Goodman's New Riddle of Induction? Philosophia, 52, 915–934.

DOI. https://doi.org/10.1007/s11406-024-00744-2

Authorship. Sole-authored by Sebastian Scholz.

Philosophia (2024) 52:915–934
https://doi.org/10.1007/s11406-024-00744-2

Conceptual Spaces: A Solution to Goodman's New Riddle of Induction?

Sebastian Scholz[1]

Received: 30 November 2023 / Revised: 23 April 2024 / Accepted: 8 May 2024 / Published online: 23 July 2024

Abstract
Nelson Goodman observed that we use only certain 'good' (viz. projectible) predicates during reasoning, with no obvious demarcation criterion in sight to distinguish them from the bad and gruesome ones. This apparent arbitrariness undermines the justifiability of our reasoning practices. Inspired by Quine's 1969 paper on Natural Kinds, Peter Gärdenfors proposes a cognitive criterion based on his theory of Conceptual Spaces (CS). He argues the good predicates are those referring to natural concepts, and that we can capture naturalness in terms of similarity. In contrast to Quine, he does not primarily rely on logic, but geometry. He frames his account as a descriptive project, however, and it is not obvious how it addresses the normative dimension of Goodman's Riddle. This paper develops a charitable reconstruction of his argument, based on the idea that the instrumental success of our projectible concepts is grounded in their cognitive-pragmatic naturalness. It then explores three lines of reasoning against the argument: Evolutionarily motivated skepticism, the miracles argument, and the relation between instrumental and pragmatic success. I conclude that in its current form, the CS account fails to provide any justification of reasoning beyond appealing to its instrumental success, and that a metaphysically robust notion of naturalness helps to achieve the desired goal.

Keywords Concepts · Induction · Naturalness · Reasoning · Scientific realism

1 Introduction

Philosophy is riddled with perplexing conundrums, and Goodman's Riddle is a particularly intriguing specimen. Goodman had noticed that we use only certain 'good' (viz. *projectible*) predicates during reasoning, with no obvious demarcation criterion

✉ Sebastian Scholz
S.Scholz@hhu.de

[1] Department of Philosophy, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

in sight to distinguish them from the bad and gruesome ones. This apparent arbitrariness undermines the justifiability of our reasoning practices – which feels important, as reasoning is generally seen as a big part of what makes us human.

Since Goodman posed his riddle in 1955, many solutions have been advanced, but the issue has proven persistent. A notably inventive approach comes from recent cognitive research. Inspired by W. O. Quine's, 1969 paper on *Natural Kinds*, Peter Gärdenfors argues the good predicates are those referring to natural concepts, and that we can capture naturalness in terms of similarity. In contrast to Quine, however, he does not primarily rely on logic, but *geometry*. His innovation is to represent similarity as a function of distance in so-called Conceptual Spaces (CS). As similar objects form clusters in these peculiar theoretical entities, properties and concepts can be defined as spatial regions – and regarded as *natural* if they display a degree of topological cohesion, as captured by the geometric property of *convexity*. The resulting demarcation criterion is praised by the author as a comprehensive solution to the riddle. But does the approach deliver what is promised? Depending on what we demand of a solution, it might not.

As a cognitive scientist, Gärdenfors is primarily concerned with explaining and predicting the behavior of cognitive agents, as well as constructing artificial agents – so he frames his account as a descriptive project. Goodman's Riddle, however, clearly has a normative dimension: It is about the *justification* of reasoning. This raises the questions of whether and how the CS account addresses the normative dimension of the riddle. In what follows, I pursue two goals. First, I will develop a charitable reconstruction of Gärdenfors' argument and thus constructively contribute to the improvement of the approach. I think the approach *does* have a clear position on the justification issue, but it is somewhat difficult to pin down and should be made more explicit. Second, I will explore three lines of reasoning that might be raised against the argument, two of which I maintain are sound. My conclusion is the approach does *not* provide justification – beyond appealing to the instrumental success of inference, which already Hume or Goodman did. In a nutshell, this is because Gärdenfors relies on a cognitive notion of naturalness, although we need ontological naturalness to make sense of the (presumed) fact that geometrically cohesive concepts are instrumentally successful in reasoning. My argument is not about convexity specifically, but it targets cognitive criteria of naturalness in general. I advocate for a conception of naturalness that takes both cognitive and ontological aspects into account.

The next sections provide background on Goodman's Riddle and what we might count as a solution. I will then introduce the CS account and work out the solution it proposes, before developing my own argument in three steps.

2 The Riddle

Goodman's New Riddle of Induction is a successor of – though logically independent from – Hume's old Problem of Induction. Hume had famously argued we base our inductive generalizations not on reason, but habit. Whenever we try coming up with reasons to justify induction, circularity looms, as we have to presuppose the unifor-

mity of past and future cases we are trying to establish.[1] Goodman extended the scope of the problem by noting that "some regularities do and some do not establish such habits" (Goodman, 1983: 82). Say, we find a bunch of green emeralds inductively confirming the hypothesis that *all emeralds are green*. This regularity and generalization habit are familiar and very intuitive. The trouble is we can formulate an alternative hypothesis, namely that *all emeralds are grue* – where *grue* is defined to be satisfied by the *very same* emeralds as follows: *green and observed before* t, *or blue and unobserved before* t. What we have before us, then, are actually at least two regularities and generalization habits. Crucially, both habits have equal evidential support, yet we prefer one over the other. How could we ever justify settling on a specific hypothesis, if we can always come up with alternative hypotheses that have equal degree of confirmation? This is the heart of the riddle and the very reason why it had such a big impact in the philosophy of science. Goodman had informally proved that the standard *instantial model of confirmation*, according to which "positive instance of a generalization lends some support to that generalization […]" (Slater & Borghini, 2013: 5) is at best incomplete.

The property of hypotheses to be confirmable by their instances, or of predicates to be transposable from one set of cases to another, is what Goodman calls "projectibility". "Green" is supposed to be projectible, while "grue" is not. As I take reasoning to be a cognitive rather than purely linguistic enterprise, I will sometimes speak more loosely of projectible concepts, but the idea is the same.[2]

2.1 What Counts as Solution?

Let us now sharpen our intuitions on what we might count as a solution. It is quite clear that a mere description of our inferential practices is not what we are after. Nobody doubts that, as a matter of fact, we make certain inductions, or project certain predicates. If, by contrast, somebody came up with robust, non-circular reasoning to justify our practices – viz. the 'strong' justification Hume was after – then this would quite certainly constitute a solution. There is, however, a lot of middle ground to cover, and the boundaries between description and prescription are sometimes muddy. Consider the following passage about Goodman's own take on the issue:

> Goodman's solution to the new riddle of induction resembles Hume's solution in an important way. Instead of providing a theory that would ultimately justify our choice of predicates for induction, he develops a theory that provides an account of how we in fact choose predicates for induction and projection. […] [He] makes projectibility essentially a matter of what language we use and have used to describe and predict the behaviour of our world. (Cohnitz & Rossberg, 2020: 5.4).

[1] Hume's argument is more intricate, of course; it takes the form of a dilemma, and circularity is the crux of one of its horns. There are also interpretations that do not make reference to the aforementioned uniformity principle; here, however, the condensed portrayal suffices. Cf. Henderson, 2020.

[2] I understand concepts here as mental representations that are postulated within cognitive psychology to explain higher cognitive competences such as categorization and reasoning.

First, note that the authors regard both Hume's and Goodman's approaches as solutions. Second, the quote could be read as if both philosophers deemed the normative dimensions of their riddles entirely irresolvable. But things are not as straightforward: Just because there is no "ultimate" justification does not mean there is none at all. Hume, for instance, was historically interpreted as a skeptic about the possibility of justification. He introduced his notion of habit as a descriptive explanation of our inductive practices, which is more than a mere description, but does not by itself address the normative question. And yet, not even Hume recommended against relying on induction in day-to-day life – so he had to think some 'weak' form of justification was available, even if just in terms of practical, everyday *success* (cf. Henderson, 2020: 5.1).

With Goodman, the idea that some of our projections are successful even takes center stage. His own answer, which is subject to the above quote, is *entrenchment*: As some of our past projections happened to be successful, the predicates and hypotheses that were involved stuck around and became entrenched in language. Or as Quine puts it: "In induction nothing succeeds like success." (Quine, 1969: 129). The notion of success provides weak justification and makes normativity part of the equation, so to speak. One should not lose sight of the fact, however, that the validity of the norms determining what counts as success cannot be derived from the description of our reasoning practices alone. Instead, for Goodman, practice itself plays a role in this, resulting in a complex overall picture that is reminiscent of the late Wittgenstein:

> What we have in Goodman's view, as, perhaps, in Wittgenstein's, are practices, which are right or wrong depending on how they square with our standards. And our Standards are right or wrong depending on how they square with our practices. This is a circle, or better a spiral, but one that Goodman […] regards as virtuous. (Putnam, 1983: ix).

Neither the details of Goodman's approach, nor the extensive criticisms that have been leveled at it, are of concern here. What all this goes to show is the title question of this paper does not have a categorical answer – we must conditionalize on what we demand of a solution. It is reasonable to require that it should somehow address the normative concern at the heart of the riddle, but if strong justification is believed to be unavailable, then finding necessary and sufficient conditions of projectibility that capture our practices may suffice. Much of the literature on projectibility can be read along these lines,[3] and Gärdenfors' convexity criterion, too, could be interpreted accordingly. I do not think, however, that this would do justice to the approach, as it tries to achieve more by invoking *naturalness*.

[3] Earman, 1985 is a good example, who formally distinguishes several problems of induction (only some of which involve gruesome predicates) and discusses candidate necessary and sufficient criteria.

2.2 The Connection with Naturalness

Gruesome predicates are weird and unwelcome. Philosophers like to label them "unnatural", "miscellaneous", or "gerrymandered", and have tried casting them out to prevent them from causing trouble (this is of course a caricature). There is no straightforward way to achieve this, however, because as Gärdenfors, 1990 points out, from a purely logical standpoint, a predicate is a predicate is a predicate. Appealing to the simplicity of the more welcome ones does not help either, because what counts as simple depends on the language in which the predicates are defined. One obvious idea is to appeal to the referents of the predicates instead. Suitable candidates include properties or kinds – however, there may be many more of those around than there are projectible predicates (cf. Lewis, 1983). According to *class nominalism*, for instance, properties are sets of individuals, so every random (or gruesome) combination of things will count as a property – which certainly doesn't help with the riddle. Platonists, by contrast, can at least in theory restrain the number of referents by appealing to *universals* (i.e., multiply instantiable entities). But anyone who wants to be metaphysically thriftier, or who looks for some practically applicable and not just theoretical demarcation criterion, is bound to have a hard time sorting this out – as does W. O. Quine, whose austerely naturalistic ontology refrains from universals and *abstract objects* (i.e., spatiotemporally distributed entities), but allows for physical objects and sets because they are indispensable for science and mathematics (cf. Hylton & Kemp, 2023).

In his influential paper on natural kinds, Quine attempts demarcation via *similarity*, the idea being that "[t]wo green emeralds are more similar than two grue ones would be if only one of the grue ones were green." (Quine, 1969: 116). Similarity, in turn, he takes to be closely related to the notion of natural kinds, even going as far as to say they are "substantially one notion" (Quine, 1969: 119). In other words, he wants to delineate among all the groupings of things an elite set of *natural* groupings of *similar* things which are depicted by the projectible predicates. Traditionally, however, naturalness is about the world having an *objective* or *mind-independent* structure for us to "carve" with our everyday or scientific "conceptual cutlery" (Slater & Borghini, 2013: 25). Thus, natural entities are usually interpreted realistically: They have been conceived of as Platonic universals, Aristotelian essences, homeostatic property clusters (Boyd, 1999), perfectly natural properties (Lewis, 1983, 1986), and many more. But due to his austere naturalism – plus his pragmatic leanings, which we will hear more about later – for Quine, these options are (or would be) out of the question. Instead, he attempts analyzing naturalness in terms of similarity, and vice versa, while also relating both concepts to more readily understandable logical or set-theoretical terms. But while he produces many valuable insights, especially into evolutionary aspects of the issue (again, more on that later), he does not find a satisfying analysis, noting that "[…] there is something logically repugnant about [similarity]" (Quine, 1969: 117) – and naturalness, for that matter.

At this point, several remarks on terminology are overdue. First, depending on the philosophical task at hand, cleanly distinguishing between natural kinds, properties, propositions, etc. is mandatory. Here this is less important, as I am primarily concerned with the contrast between *cognitive* and *ontological* naturalness of any

kinds of entities. I will thus sometimes speak interchangeably of natural *concepts* or *representations* to indicate the cognitive notion, versus natural *kinds* or *properties* to indicate the ontological notion. The nature of this distinction will be elaborated in Sect. 3.2. It should be noted that Gärdenfors treats properties as cognitive entities, which is a source of potential terminological confusion that I will do my best do avoid.

Second, I will reserve the term *strong realism* for the ontologically committed view that there are mind-independent entities which are the natural kinds. This is the position I will argue for; it gives the natural entities an *intrinsic* metaphysical status. According to Bird & Tobin, *weak realism* is by contrast the "ontologically uncommitted view that our classifications are often natural." (Bird & Tobin, 2023: 1.1.1). Yet, in this discussion, I will include positions such as Quine's, which affirm the reality and objectivity of the kinds, but do not assign them an intrinsic metaphysical status. Quine instead highlights their dependence on our scientific interests. Given that this stance incorporates elements traditionally associated with *conventionalism*, which is "the view natural kinds don't exist independently of the scientists and others who talk about them" (Bird & Tobin, 2023: 1.1.2), it can be misleading to classify him as a strong realist, despite the technical possibility of doing so.

Now, according to Judith Crane, "[p]hilosophical treatments of natural kinds are embedded in two distinct projects. […] The kinds studied in the philosophy of science approach are projectible categories that can ground inductive inferences and scientific explanation. The kinds studied in the philosophy of language approach are the referential objects of a special linguistic category—natural kind terms—thought to refer directly." (Crane, 2021: 12177). Our present concern is the former project, and although Crane does not mention him explicitly, we may think of Quine as one of its key historical figures. Couched in modern terminology, thus, his 1969 paper is about *grounding* projectibility in naturalness – although he was ontologically more parsimonious than many recent grounding enthusiasts, so this is to use 'grounding' in a broad sense akin to 'metaphysical explanation'. In order not to open a big can of worms, I will not get into the details of the grounding debate. However, there is an important line of thought for us to distill here. Grounding reasoning in the structure of the natural world comes down to at least two things: Taking the latter to be more fundamental than the former (cf. Fine, 2001 on the relation between grounding and fundamentality), and, crucially, taking the former to be justifiable insofar as it corresponds to the latter. This second point is perhaps so banal that it is rarely stated explicitly, but it is implicit in commonplace talk about how natural kinds "enable us" (Khalidi, 2013: 72) to make inductive inferences, for an instance. The justification strategy to derive here is that, simply put, *if we can provide a sound metaphysical explanation for why some predicates are projectible, then we have all the more reason to trust them*. Whether this approach can generate anything resembling a 'strong' justification is unclear and will highly depend on one's meta-metaphysical views, but it is certainly more ambitious than the familiar appeal to instrumental success. And I think this is precisely the move Gärdenfors is trying to make along the normative dimension of the riddle.

3 Conceptual Spaces

Gärdenfors' diagnosis of Quines 'failed' analysis of naturalness is that he relied too heavily on the methodology of the Vienna Circle: "Using logical analysis, the prime tool of positivism, is of no avail […]. [We] have to go below language." (Gärdenfors 2011: 3). Instead of focusing on the symbolic, or propositional form of inference, he suggests looking into how we represent the conceptual knowledge that is involved. And the format of our conceptual representations, he contends, is *geometric* in nature.[4] It is this move that allows him to treat similarity as spatial distance, namely as distance in CS.

Conceptual Spaces are cognitive and mathematical theoretical entities. Cognitive, as they are postulated to explain and predict the behavior of cognitive agents, and mathematical, as they are based on axiomatic geometry and usually come with metrics (Eucledian or other, the details do not matter here). Moreover, CS typically come with *quality dimensions*, whose "primary function […] is to represent various 'qualities' of objects." (Gärdenfors, 2009: 5). Dimensions can be *separable* or *integral*: The visual color domain, which is Gärdenfors' paradigm case, consists of the dimensions *hue*, *saturation*, and *brightness*. Possible or real visual objects, as represented by points in the space, cannot have a value in one of these dimensions without having one in the others, and in this sense the latter are inseparable, viz. integral. With this conceptual apparatus in place, *properties* are defined as regions in CS-domains, *concepts* as regions across multiple domains. The property of being green, for instance, is but a region in the visual domain.

3.1 Proposed Solution

To see how this might help with the riddle, recall Quine's take on projectibility. The upshot was projectible predicates are those referring to natural classes of similar entities. I just outlined how Gärdenfors translates similarity into spatial proximity. As a result, similar objects belonging to a natural class will form clusters in CS, the degree of cohesion of which can be taken to indicate the degree of naturalness of the respective classes. Gärdenfors discusses several topological cohesion criteria, but the one he settles on is *convexity*: A region is convex, iff for any points A and B within the region, there is no point C between A and B that is not in the region as well. The resulting definitions read as follows:[5]

(P) A *natural property* is a convex region in some domain.
(C) A *natural concept* is represented as a set of convex regions in a number of domains together with a prominence assignment to the domains and information about how the regions in different domains are correlated.

[4] It must be noted that the CS approach leaves room for other, symbolic or sub-symbolic representational formats, so it is actually pluralistic about the structure of mental content.

[5] Gärdenfors employs slightly differing definitions, these are taken from (2008).

To sum up the proposal in a nutshell: Convexity figures as a demarcation criterion for projectible predicates via demarcation of the natural representations which are their referents.

Two things are important to note: First, the connection between convexity and naturalness is meant to be empirical, not analytical. If some other cognitive criterion would fit the data better, Gärdenfors would presumably go with that, without his superordinate account being in jeopardy. Second, convexity is meant to be a necessary, but not sufficient criterion. Being non-black, for instance, is a convex region of the visual space, but does not feel very natural as a property (the example is taken from Hempel's Paradox of Confirmation, which is the other riddle Quine discusses). But for Goodman's case, at least, the criterion seems to get rid of the undesirable predicate: Being green is convex and satisfies the necessary condition of naturalness. Being grue, by contrast, is non-convex in a visual space plus time dimension, as the corresponding region disconnects at time=t. There is some discussion on whether this is convincing (cf. Hernández-Conde, 2017, or Strößner 2022), however, my argument will not depend on the intelligibility of convexity specifically, but apply to cognitive naturalness criteria more generally, so there is no need to engage in it here.

Let me now reconstruct Gärdenfors' argument to the conclusion that convexity in CS is the solution to Goodman's Riddle. As was previously mentioned, CS are largely about explaining and predicting behavior. Since projecting some predicates but not others is part of our 'inferential behavior' as cognitive agents, convexity is supposed to account for that. A genuinely descriptive premise can be derived, analogous to Hume's notion of habit:

P1. Our most projectible concepts are convex (cognitively-descriptively)

[6]Since the issue expressed in P1 is an empirical rather than a conceptual matter, I will not try resolving it here. The same is true of the second premise, which – analogously to Goodman's notion of entrenchment – is about instrumental success:

P2. Convex concepts have been instrumentally most successful in reasoning

[7]This premise is still descriptive, but it makes normativity part of the equation, as success is a normative notion. Why presume Gärdenfors would subscribe to P2? On a general level, he repeatedly emphasizes his affiliation with *scientific instrumentalism*, the view that scientific theories are tools for predicting and controlling experiences rather than true descriptions of reality. But you can also see that by the fact that he wants to apply convexity to build reasoning AI, so he has to think the approach leads to somewhat successful reasoning.

[6] P1 speaks of 'concepts' instead of 'predicates' to indicate Gärdenfors' focus on cognition over language. Another way to express this would be to say that our most projectible predicates *refer* to convex concepts. Within *cognitive semantics*, which feeds into the theoretical background of the CS account, linguistic expressions are generally thought to refer to cognitive entities.

[7] If a solution to Hume's Problem of Induction is available, then this premise can be generalized to future cases. This is supposed to be a circumspect reconstruction of Gärdenfors' argument, however, so a less assumptive formulation was chosen.

Note that P2 can be considered an *explanans* for P1 as *explanandum*: It is *because* of their instrumental success that convex concepts are projectible. Now, the way I think CS take things a step further is by invoking naturalness as an *explanans* for P2 as *explanandum*:

P3. P2 holds because of the convex concepts' naturalness.

This is the crucial step described above which amounts to grounding instrumental success in naturalness to achieve better justification: Convex concepts are reliable, because they are natural, and that's why we should trust them. This is Gärdenfors' answer to the normative part of the riddle, which has so far only been implicit in his writings.

The success of this argument will of course heavily depend on how one conceives of naturalness, and it is here where my critique will hopefully have some bite. I also owe some substantiation of the sentiment that Gärdenfors would buy into a version of P3. At any rate, it is required to delve a bit deeper into the metaphysical and epistemological intricacies of the approach.

3.2 Metaphysical and Epistemological Ramifications

As previously stated, most traditional theories of naturalness treat natural entities as mind-independent and objective, i.e., they advocate for some version of strong realism. Gärdenfors, however, attacks mind-independence:

> […] when it comes to inductive *inferences* it is not sufficient that the properties exist out there somewhere, but we need to be able to grasp the natural kinds by our minds. In other words, what is needed to understand induction, as performed by humans, is a *conceptualistic* or *cognitive* analysis of natural properties. (Gärdenfors 2011: 3).

But cognitive analyses of this sort are sometimes explicitly dismissed by contemporary metaphysicians such as David Lewis, whose writings continue to have considerable influence on the naturalness debate:

> Nor should it be said […] that as a contingent psychological fact we turn out to have states whose content involves some properties rather than others, and that is what makes it so that the former properties are more natural. (This would be a psychologistic theory of naturalness). (Lewis, 1983: 377).

We are thus dealing with two distinct notions of naturalness – a *cognitive* and an *ontological* one. I will take ontological naturalness to mean strong realism, although nominalist accounts could be accommodated as well. The important point here is naturalness depends on the structure of mind-independent reality.[8] Cognitive natural-

[8] The criterion of mind-independence is a *proviso* and might turn out to be problematic (cf. Khalidi, 2016). I am confident that there are other (non-cognitive) ways of spelling out ontological naturalness (e.g., in

ness, by contrast, first appears to be a type of full-blown conventionalism. However, as will be explored throughout this section, one can coherently devise a cognitive analysis of naturalness without giving up the notion that the world plays *some* role in shaping our natural concepts, i.e., cognitive naturalness does not necessitate *strong* conventionalism. Such analysis is precisely what Gärdenfors aims for when he says it is not sufficient that the properties exist out there – but does not rule out that they do.

But how exactly is this supposed to work? Does that mean his approach is an intermediate position between conventionalism and realism, and if so, doesn't it run the risk of collapsing into one side or the other? This point is important and deserves some scrutiny. It runs parallel to the debate around *perspectival realism*, where some philosophers try to account for the perspectival nature of, say, how we theorize about kinds, but without giving up on non-perspectival aspects, viz. realism – which has proven difficult to achieve. This is a remarkable cross-connection I can merely allude to here.[9] To see how the CS account navigates the issue, let us first get a clear picture of its conventionalist leanings.

Criteria (P) and (C) tie naturalness to convexity, i.e., to the structure of our cognitive representations. It is fair to say, then, that this characterization *localizes* naturalness in the heads of cognitive agents. And indeed, Gärdenfors explicitly endorses the affirmative version of Hilary Putnam's famous slogan "[…] meanings just ain't in the head!" (Putnam, 1975a: 227), that is he endorses *semantic internalism*. He does make an adjustment to account for the social dimension of meaning, as indicated by the plural form "heads" (cf. Gärdenfors, 1999a), but this appears to fit neatly with the view that meanings – including those of natural kind terms – are a matter of (social) convention and have nothing to do with mind-independent facts. One of the hallmarks of semantic internalism is the notion that *intensions fix extensions* (cf. Schrenk, 2016: 238–39). Within philosophy of science, this view can amount to the meanings of theoretical terms (including natural kind terms) being wholly determined by their respective background theories, which implies what is known as *Kuhnian relativism*.[10] And indeed, Gärdenfors is compelled to address this very issue:

> But is not the choice of a conceptual space arbitrary? Since a conceptual space may seem like a Kuhnian paradigm, aren't we thereby stuck with an unavoidable relativism? After all, anyone can pick her own conceptual space and in this way make her favorite properties come out natural in that space. (Gärdenfors, 2000, Sect. 3.7).

Let us consider for a moment where the quality dimensions of CS come from: They are not given by God, but chosen by scientists who try to model cognition. They are

terms of fundamentality), however, this needs to be done in a different paper.

[9] There have been attempts to spell out the notion of scientific perspective within the CS framework, e.g. Kaipainen & Hautamäki 2015. The authors speak of perspectives on so-called "ontospaces", but that does not seem to involve a strong metaphysical commitment.

[10] Kuhnian relativism makes truth, knowledge, or meaning relative to Kuhnian *paradigms* (cf. Kusch, 2021). Thomas Kuhn introduced the latter concept in his *Structure of Scientific Revolutions* as "universally recognized scientific achievements that for a time provide model problems and solutions to a community of practicioners." (Kuhn, 2012: *Preface*).

often the result of mathematical regression analyses (e.g. multi-dimensional scaling) being applied to psychological data such as similarity judgements, aiming for a high fit between judgements and spatial distances in the model. If criteria P and C tie naturalness – and thus projectibility – to these models, doesn't that introduce precisely the kind of capriciousness you would expect of epistemic relativism? Gärdenfors speaks of a "[…] *metaproblem* of inductive inferences: What criteria can be used to choose between competing conceptual spaces?" (Gärdenfors, 1990: 94).

This is the nature of the challenge, and it might be answered as follows. First, it should be noted that internalism about linguistic meaning must be distinguished from internalism about mental content. One can consistently tie the meanings of natural kind terms to exclusively their intensions without claiming the same for their mental counterparts – which is precisely the path Gärdenfors seems to take by asking "*Does Semantics Need Reality?"*, and replying:

> […] "not directly." Once we accept the conceptual structure of an individual as given, the semantic mapping […] can be described without any recourse to the external world. But a second part of the cognitivistic answer is "indirectly," since the conceptual structure is built up in an individual in interaction with reality. (Gärdenfors, 1999b: 14).

Moreover, recall that convexity is intended only as a necessary naturalness criterion. That is why he can claim his analysis "[…] is compatible with, though does not require, strong metaphysical realism." (Gärdenfors, 1990: 90).

One may wonder, however, whether this is true and whether he does need some version of strong realism in order to avoid relativism, after all. Partly, this concern is addressed by appealing to coherence (instead of convention) with an "extended net of empirical knowledge" (Mormann 1993: 236) – which certainly makes things less arbitrary, but might still be compatible with relativism. The more important argument is once more inspired by Quine – who has phrased the key idea as poignantly as it gets: "Creatures inveterately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing their kind." (1969, p. 126). The reason we can trust our inferences is natural selection would have wiped us out a long time ago if we could not. By applying this line of thought to our internal cognitive make-up, Gärdenfors installs evolution as a 'bridgehead' between cognition and world. He fills this general point with life by emphasizing the cognitive *efficiency* of convex concepts (cf. Douven & Gärdenfors 2019 for elaboration of this point). With respect to the metaphysics of naturalness, this leaves us with weak realism that is enriched with *pragmatic* elements:

> Via successful and less successful interactions with the world, the conceptual structure of an individual will adapt to the structure of reality. It must be emphasized, however, that this does not entail that the conceptual structure *represents* the world. (Gärdenfors, 2000: 156).

The bottom line is this: While the world *does* shape our cognition, we cannot say much at all about *its* structure. Our relation to it is not conceived of in terms of truth

or correspondence, as would be typical for strong realism, but as a matter of practical problem solving. It is basically viewed as a Kantian thing-in-itself, and indeed Gärdenfors labels his epistemological position as neo-Kantian (cf. Gärdenfors, 2000: fn. 161, 171). A more precise version of P3 can now be derived:

$P3_{CP}$. P2 holds because of the convex concepts' *cognitive-pragmatic* naturalness.

By now it should be clear that P1 and P2 alone are not sufficient for a charitable reconstruction of the argument. Naturalness is meant to contribute decisively to the avoidance of relativism, which distinguishes the proposal from Goodman's (in the normative dimension, that is, differences in their descriptive accounts should be obvious). $P3_{CP}$ adequately captures the spirit of this approach.

To sum up, the metaphysics and epistemology of naturalness in CS build upon Quine's take on the riddle. Natural entities are thought to be real and objective, but still dependent on what practical use they have for us. While Quine tends to emphasize the roles of science and language, Gärdenfors highlights the role of knowledge representation. Cognitive naturalness certainly has a high affinity to conventionalism, but within the framework of Quinean pragmatism a weakly realistic version of the notion can be obtained.

4 The Case for Strong Realism

The remainder of the paper will investigate three lines of reasoning that can be advanced against the argument I just presented. One is as obvious as it is interesting, but it undermines my own position as well – fortunately, it can be refuted. The second one argues positively in favor of strong realism. The third one expresses a concern, rather than providing a knock-down argument. All three points extend into several major debates, so what follows is an explorative rather than an exhaustive discussion. Let me go through them one by one.

4.1 Adaptivity of Poor Reasoning

The theory of evolution is powerful, but one must always be careful not to strain it. It is a common mistake to assume that evolutionary processes automatically lead to progress, and perhaps we are dealing here with an epistemological instance of this fallacy. Gärdenfors and Quine seem to put a lot of hope into evolution when they effectively claim the adaption of cognitive agents to their environment leads to reliable inductive reasoning. But as has been argued many times (cf. Stich, 1990: 55–74 for a systematic philosophical discussion): *unreliable reasoning can be highly adaptive.* Just think of the rich psychological literature on the countless ways in which human cognition has evolved to be biased. If naturalness is tied to what facilitates survival and reproduction, maybe we should not expect it to remedy Goodman's Riddle.

Unfortunately, this point does not only target $P3_{CP}$, but also the brand of scientific realism I want to advocate here. If the evolutionary perspective gives us reason

distrust of our cognitive faculties, then we surely must also doubt we have reliable access to mind-independent reality: "[…] a fancier style of representing is advantageous *so long as it is geared to the organism's way of life and enhances the organism's chances of survival*. Truth, whatever that is, definitely takes the hindmost." (Churchland, 1987: 549).[11] Or as Gärdenfors might frame the issue: Organisms adapt to their cognitively reshaped, *proximal* environments, which they encounter in the form of practical problems (food retrieval, etc.). They have no access to a common *distal* environment, which thus takes on the status of a Kantian Thing-in-itself.[12] There may be an insurmountable gap between what is evolutionarily useful and what the world is really like, call this the evolutionary argument against realism, or EAAR.[13]

Several countering strategies are on the market, but I will focus on a response that would be natural for the CS account to give and then make my way to the second argument against $P3_{CP}$ – which happens to be a cardinal argument against EAAR. Both points I am going to make have to do with one of my favorite insights, namely that *science works*. The first is more direct, countering the claim of unreliability. The second is more ambitious, trying to account for why our faculties might be reliable by suggesting that they capture the structure of the world.

Now, the approach to take is to first point out that even if much of our 'primordial' reasoning is flawed, we have developed means to correct for these flaws. Couched in Gärdenfors' terminology, one could say that some of our *phenomenal* (i.e., subjective, derived from psychological data) CS may be prone to error, but many of our *scientific* (i.e., idealized, objective) spaces are more reliable.

> As Quine (1969) notes, it is a sign of mature science that notions of similarity become less and less important, being replaced by theoretically more sophisticated concepts. […] In this way science builds upon our more or less evolutionarily determined conceptual spaces, but in its most mature form becomes independent of them. (Gärdenfors, 2000: 83).

We are no longer deceived by superficial similarities of dolphins and fish, for instance, but figure the former are actually mammals. This rejoinder appeals to the instrumental success of science and restores hope in the pragmatic utility of our natural concepts: They are at least reliable enough to provide a basis for science to operate on. Note, however, that this move turns Gärdenfors' argument – as reconstructed above anyway – on its head. All the emphasis is on instrumental success now, which was

[11] This is not to say Churchland is a scientific anti-realist, she just exemplifies this line of reasoning nicely.

[12] Gärdenfors, 2000 cites Thompson, 1995 on this, who develops an empirical argument against the (teleofunctional) supposition of distal representational objects: "The distal properties detected in color vision form a heterogeneous collection whose type-divisions at the physical level do not match the type-divisions at the perceptual level." (Thompson, 1995: 6).

[13] Plantinga 1993 develops an evolutionary argument against naturalism (EAAN) based on the idea that naturalists should believe in evolution, and evolution undermines the reliability of our beliefs – including our beliefs in naturalism and evolution (see Beilby, 2002 for a collection of essays on the issue). I do not want to invoke the details of his argument, however, but highlight the general line of thought is well-known in the philosophical literature.

supposed to be grounded by naturalness. Against this background, the chances of providing more than an instrumentalist justification of reasoning seem dim.

Let me illustrate this line of thought with an example. *Animism* is roughly the tendency to attribute agency to inanimate objects. From the perspective of evolutionary psychology, it can be seen as a reasoning bias that evolved "[…] because those interpretive bets that aim highest (by attributing the most organization and hence significance to things and events) have the greatest potential payoffs and lowest risks. For example, it is better for a hiker to mistake a boulder for a bear than to mistake a bear for a boulder." Guthrie, 1993: 6. It is therefore plausible to assume that evolution has programmed us to systematically misjudge the degree of organization of our environment, which can lead to generalized doubts about our epistemic and reasoning capacities. To counter this, instrumentalists can point out that science enables us to study biases such as animism, which may put us in a position to avoid them. Moreover, on a general level, science often works remarkably well, which also speaks against generalized skepticism. This countering strategy is also available to Gärdenfors, but adopting it would suggest tying the justification of scientific reasoning to purely instrumental considerations, after all. Appealing to cognitive naturalness is not an option here, as this is the very notion that invites skepticism in the first place. But it would feel contrived to involve naturalness in the discussion on Goodman's Riddle, but not in a reply to evolutionarily motivated skepticism, as both undermine trust in our reasoning capacities. This suggests that Gärdenfors can either bite the bullet of evolutionarily motivated skepticism, or his theory of naturalness collapses into instrumentalism.

4.2 No (Naturalness-) Miracles

I suggest to improve on the account by making naturalness more metaphysically robust. Yes, many of our natural concepts and scientific generalizations *are* reliable – and this is because they capture nature's structure. Reasoning that attributes agency to inanimate objects is flawed precisely because it does not capture the relevant structure. Ghosts and their ilk are not natural kinds, even if they are represented by convex CS regions – which they presumably are, by the way. While it is at least feasible we've got everything wrong, this seems unlikely, given some of us can send rockets to Mars, vaccinate against Polio, and so on. We simply unveiled enough of the world to achieve these successes – at times *despite* evolution having shaped our cognition. It is only with real natural kinds, properties, etc. in place that the reliability of reasoning makes sense in the first place. This is not engaging in speculative metaphysics, mind you, but a scientifically informed *inference to the best explanation*, or IBE argument.

While this approach is metaphysically more demanding than Quine's, it is meant to be aligned with his project of *naturalized epistemology*. According to Kornblith (1993), this project is "[…] addressed to two questions: (1) What is the world that we may know it?; and (2) What are we that we may know the world?" Crucially, both answers "must dovetail in important ways" (Kornblith, 1993: 2), and they must be given without recurse to any *prior philosophy* (cf. Hylton & Kemp, 2023: 5). The latter comes down to taking science seriously, and concerning question (2), cognitive

psychology tells us that our access to the external world is *mediated* by the concepts we categorize and reason with. Even if our scientific representations become ever more idealized, objective, etc. – we cannot attain the infamous god's eye view. Therefore, I do not advocate reducing the issue of naturalness to its ontological dimension, viz. to question (1), but opt for some explication that contains both a necessary cognitive component – whether it turns out to be convexity or anything else that is empirically robust – *and* a necessary ontological component.[14] For the CS account of Goodman's Riddle, this would mean to replace $P3_{CP}$ with:

> $P3_{CO}$ P2 holds because of the convex concepts' *cognitive-ontological* naturalness.

Prima facie, the ontological part has no *practical* consequences for *descriptively* distinguishing natural from non-natural concepts, or projectible from non-projectible predicates. For that matter, Gärdenfors would likely insist "that adding a metaphysical component (the natural kinds *an sich*) to the conceptual spaces does not improve our understanding of natural kinds." (Gärdenfors, 1990: 90). But it does have important *theoretical* implications for the justification of our projections, as it makes sense of the very fact that some of our concepts have been reliable to project. It is not enough to point to the evolutionary utility, cognitive efficiency, etc. of the concepts, as their utility and efficiency are revealed at the inescapable touchstone of reality. There is, however, a remarkable paper where Gärdenfors appears to share this view.

But let me take a quick step back. The above argument is, of course, a reiteration of the famous no miracles argument, according to which realism "is the only philosophy that doesn't make the success of science a miracle." (Putnam, 1975b: 73). It's strengths and weaknesses are well known – for instance, that it expresses a powerful intuition shared by many philosophers.[15] Or, on the other hand, that there are valid concerns regarding the so-called *base rate fallacy* (cf. Chakravartty, 2017: 2.1). I will not get into the debate here, but what I want to point out is Kornblith (1993) makes an even stronger claim than Putnam does. He develops a *transcendental* argument based on the *modal* claim that reliable induction is impossible without the world playing its part – and curiously enough, Gärdenfors & Stephens, 2017 quote him on this:

"We might not have unbiased contact with the world, but it is still the real world that provides the sensory input we get, and "[i]t is precisely because the world has the causal structure required for the existence of natural kinds that inductive knowledge

[14] I personally align with the views of Kornblith (1993) (except for his hardcore essentialism) and support Boyd's Homeostatic Property Cluster (HPC) account, which emphasizes accommodation to causal structure. Boyd's stance on the metaphysics of natural kinds is nuanced: He attributes an intrinsic metaphysical status to the property clusters, yet considers kinds to be real only to the extent that they meet the accommodation demands of *disciplinary matrices*, i.e. collective cognitive and methodological environments in which scientific inquiry and classification take place. Perhaps we can think of these cognitive environments in terms of CS – in any case, Boyd emerges as a strong realist who integrates cognitive elements into his theory.

[15] Note that the *abductive* inference that the argument makes reflects the methodology of the natural sciences. This is in keeping with Quine's naturalism, but also with the contemporary research program of *inductive metaphysics*, which emphasizes the role of both aprioristic *and* empirical methods for metaphysics (see Engelhard et al. 2021 for reference). Both feed into the methodological background of this paper.

is even possible." (Kornblith, 1993, p. 35). There are only certain clusters of properties that are organised in a stable enough way as to stick together in natural categories enabling us to make inductive inferences."

Not only is this an endorsement of strong realism, but of a quite demanding premise, as such modal claims can be difficult to substantiate. However, this stance is an utter exception in Gärdenfors' work, and can be attributed with some certainty to his collaborator.[16] Even whether the less demanding abductive inference is justified is subject to vigorous debate – however, if the premise can be made to work for scientific realism in general then it will transfer to natural kinds and their ilk, as these are the kinds of entities that are regularly postulated by the special sciences.

With that, there is one more thing I would like to add in this section. In a sense, the IBE miracles argument runs bottom to top: We start with empirical facts about creatures like us – in this case the way we make inferences with convex concepts – and end with real natural kinds as ontological overlay. We can also turn that around: Assuming that realism about kinds is true (and we are thus subject to evolution), we should expect inferentially successful and cognitively economic concepts. Both directions taken together constitute a *package deal* argument for strong realism about natural kinds.[17]

4.3 Cognitive-Pragmatic Naturalness – A Poor Explanation?

I have just presented the outlines of a positive argument for a cognitive *cum* ontological approach to naturalness. To round things off, I will briefly return to undermining $P3_{CP}$. Recall the notion of cognitive-pragmatic naturalness is supposed figure as an *explanans* for the instrumental success of convex concepts (viz. P2) as an *explanandum*. My concern is that $P3_{CP}$ might be more of a restatement than an explanation for P2. This is because one appeals to instrumental success, while the other to evolutionary success, and these two might be closely related. If they turn out to imply one another, then we have before us an instance of circular reasoning.

The point seems to hang on what is considered a success in each case. As previously stated, success is a normative notion and implies the presence of some function, or goal. A typical goal in the context of scientific instrumentalism is reliable prediction. There appear to be many cases in which this coincides with evolutionary goals such as survival and reproduction. Reliably predicting the toxicity of a mushroom should aid in survival, which could be taken to indicate $P2 \supset P3_{CP}$. However, if framed a bit differently, this same example might be used to disambiguate instrumental and evolutionary success. For instance, if a strategy of overly cautious mushroom categorization – yielding many false positives – enhances survival, it could be considered evolutionarily reliable *qua* being instrumentally unreliable, i.e., $\neg(P3_{CP} \supset P2)$.

Unfortunately, the issue is too complex to be resolved in this paper, as it would quickly devolve into a systematic comparison of scientific instrumentalism and prag-

[16] Gärdenfors has confirmed to me in a personal conversation that he has not become a strong realist.

[17] I would like to thank Markus Schrenk, who brought up the idea of a package deal in a discussion, thus complementing the IBE-argument.

matism – which is a mammoth project. But even assuming the desired disambiguation is feasible, we are still left with the diagnosis that on the CS account, grounding instrumental success in naturalness amounts to grounding it in evolutionary success. The last sections have cast doubt on whether this is a good idea. To be clear, I do not want to claim strong realism is the only game in town as a theory of naturalness. 'Cognitive-ontological naturalness' means I am open to nominalism as well as weak conventionalism, as long as a mind-independent component is involved. I also want to highlight that Gärdenfors' approach demonstrates appealing to naturalness is possible for instrumentalists, pragmatists, and the New Realists who want to get rid of the traditional commitment to mind-independence (e.g., Chang's, 2022 operational coherence, Massimi's, 2022 perspectival realism, or Kendig's, 2015 practice grounded kinds). It is an open question, however, whether these and comparable authors would subscribe to grounding reasoning in evolution. After all, even if circularity can be avoided and EAAN as well as independent arguments for strong realism can be refuted, the normative question remains addressed in a very familiar way. The quote from Sect. 2.1 by Cohnitz & Rossberg, 2020 can be slightly adjusted to illustrate this: "[Gärdenfors] makes projectibility essentially a matter of what [concepts] we use and have used to describe and predict the behaviour of our world" – not the world in and of itself though, but an evolutionarily reshaped version of it. This move may enable one to sidestep *Goodmanian* relativism, but at the cost of an evolutionary one. Whether this is much progress, I leave for the reader to decide.

5 Conclusion

Returning to the title question of this paper: Are Gärdenfors' Conceptual Spaces a viable solution to Goodman's Riddle? They most certainly have something to offer. Convexity, as a necessary condition of projectibility, leads to rich empirical predictions and – given those predictions will obtain in the long run, which I have not debated here – provides a descriptive explanation of our inferential practices, as well as a practically implementable demarcation criterion. This has to count as progress, especially from a cognitive science perspective – but it is progress in the descriptive dimension of the riddle. If progress in the normative dimension is the benchmark, then, as I have argued, CS in their current form miss it.

Section 2.1 has shown that already Hume and Goodman saw the instrumental success of some of our inferences provides a basis for their 'weak' justification. Gärdenfors can be interpreted as appealing to instrumental success as well (P2), but I have presented a reconstruction of his account that aims for a wider justification strategy by grounding instrumental success in cognitive-pragmatic naturalness ($P3_{CP}$). It is doubtful, however, whether this notion of naturalness is up to the task. Evolutionarily motivated skepticism may be avoidable, but the most obvious countering strategy appeals to the success of science (i.e., no progress in the normative dimension), as appealing to naturalness is not an option, if it is tied to evolutionary success – which is the very notion that invites skepticism in the first place. As a result, the wider justification strategy collapses into the familiar instrumentalist one. By contrast, appealing to naturalness is very straightforward on a strongly realist reading: Reasoning

is justified insofar as it captures the structure of nature. Section 4.2 has outlined a 'package deal' argument in favor of giving natural entities an intrinsic metaphysical standing – but without disregarding the progress that comes with cognitive criteria of naturalness. My positive case is partially based on the miracles argument from the great debate about scientific realism, where I'm in good company with philosophers like Hilary Putnam. My negative case against cognitive-pragmatic naturalness is strengthened by considerations about the relation of instrumental and evolutionary success, which the CS account would need to disambiguate more clearly for the grounding justification strategy to work.

I agree with Gärdenfors that going below the level of language to understand reasoning is the right move. But it is ill-advised to tie the justification of reasoning exclusively to how we represent the world around us. It is the structure of the natural world itself that determines which of our inferences work and which do not.

Acknowledgements The research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG) under project number 462340335. I would like to thank my colleagues and supervisors in Düsseldorf, especially Daian Bican, Giacomo Gianni, Jan Michel, Markus Schrenk, Gerhard Schurz, and Gottfried Vosgerau, for their support and most helpful discussions.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Beilby, J. K., Ed. (2002). *Naturalism defeated? Essays on Plantinga's evolutionary argument against naturalism*. Cornell University Press

Bird, A., & Tobin, E. (2023). 'Natural Kinds'. In E. N. Zalta, & U. Nodelman (Eds.), *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/natural-kinds/.

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays*, (pp. 141–185). MIT Press.

Chakravartty, A. (2017). Scientific realism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/.

Chang, H. (2022). *Realism for realistic people: A new pragmatist philosophy of science*. Cambridge University Press. https://doi.org/10.1017/9781108635738.

Churchland, P. S. (1987). Epistemology in the age of neuroscience. *Journal of Philosophy*, *84*(10), 544–553. https://doi.org/10.5840/jphil1987841026.

Cohnitz, D., & Rossberg, M. (2020). Nelson goodman. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy, Summer* (pp. 1–25). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/goodman/.

Crane, J. K. (2021). Two approaches to natural kinds. *Synthese*, *199*(5–6), 12177–12198. https://doi.org/10.1007/s11229-021-03328-9.
Douven, I., & Gärdenfors, P. (2019). What are natural concepts? A design perspective. *Mind and Language*, *35*(3), 313–334. https://doi.org/10.1111/mila.12240.
Earman, J. (1985). Concepts of projectibility and the problems of induction. *Noûs*, *19*(4), 521–535. https://doi.org/10.2307/2215351.
Engelhard, K., Feldbacher-Escamilla, C. J., Gebharter, A., & Seide, A. (2021). Inductive metaphysics: Editors' introduction. *Grazer Philosophische Studien*, *98*(115), January, 1–26. https://doi.org/10.1163/18756735-00000129.
Fine, K. (2001). The question of realism. *Philosophers' Imprint*, *1*, 1–30.
Gärdenfors, P. (1990). Induction, conceptual spaces and AI. *Philosophy of Science*, *57*(1), 78–95.
Gärdenfors, P. (1999a). Some tenets of cognitive semantics. In J. Allwood, & P. Gärdenfors (Eds.), *Cognitive semantics: Meaning and cognition* (pp. 19). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.55.03gar
Gärdenfors, P. (1999b). Does semantics need reality? In A. Riegler, M. Peschl, & A. Stein (Eds.), *Understanding representation in the cognitive sciences* (pp. 209–217). Springer. https://doi.org/10.1007/978-0-585-29605-0_23
Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press Cambridge, Massachusetts
Gärdenfors, P. (2008). Reasoning in conceptual spaces. *Reasoning: Studies of human inference and its foundations* (pp. 302–320). Cambridge University Press. https://doi.org/10.1017/CBO9780511814273.018.
Gärdenfors, P. (April 2009). Meanings as conceptual structures. *Lund University Cognitive Studies*, *16*, 1–13.
Gärdenfors, P. (1 January 2011). Inductive reasoning: From carnap to cognitive science. 1–16.
Gärdenfors, P., & Stephens, A. (2017). Induction and knowledge-what. *European Journal for Philosophy of Science*, *8*(3), 471–491. https://doi.org/10.1007/s13194-017-0196-y.
Goodman, N. (1983). *Fact, fiction, and forecast*. Vol. 4th Edition. Harvard University Press.
Guthrie, S. E. (1993). *Faces in the clouds: A new theory of religion*. Oxford University Press USA.
Henderson, L. (2020). The problem of induction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy,*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/induction-problem/.
Hernández-Conde, J. (2017). A case against convexity in conceptual spaces. *Synthese*, *194*(10), 4011–4037. https://doi.org/10.1007/s11229-016-1123-z.
Hylton, P., & Kemp, G. (2023). Willard van orman quine. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/quine/.
Kaipainen, M., & Hautamäki, A. (2015). A perspectivist approach to conceptual spaces. In F. Zenker and P. Gärdenfors (Ed.), *Applications of conceptual spaces: The case for geometric knowledge representation*, 245–258. Synthese Library. Springer International Publishing. https://doi.org/10.1007/978-3-319-15021-5_13.
Kendig, C. (2015). Editor's introduction: Activities of kinding in scientific practice. *Natural kinds and classification in scientific practice* (pp. 1–14). Routledge.
Khalidi, M. A. (2013). *Natural categories and human kinds: Classification in the natural and social sciences*. Cambridge University Press.
Khalidi, M. A. (2016). Mind-dependent kinds. *Journal of Social Ontology2*(21 August), 223–246. https://doi.org/10.1515/jso-2015-0045.
Kornblith, H. (1993). *Inductive inference and its natural ground: An essay in naturalistic epistemology*. The MIT Press
Kuhn, T. S. (2012). *The structure of scientific revolutions: 50th Anniversary Edition*. The University of Chicago
Kusch, M. (2021). *Relativism in the philosophy of science*. Cambridge University Press.
Lewis, D. (1983). New Work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343–377. https://doi.org/10.1080/00048408312341131.
Lewis, D. (1986). *On the plurality of worlds*. Wiley-Blackwell.
Massimi, M. (2022). *Perspectival realism. Oxford studies in philosophy of science*. Oxford University Press
Mormann, T. (1993). Natural predicates and topological structures of conceptual spaces. *Synthese*, *95*(2), 219–240. https://doi.org/10.1007/BF01064589.

Putnam, H. (1975a). *Philosophical papers: Volume 1: Mathematics, matter, and method*. Cambridge University Press

Putnam, H. (1975b). *Philosophical papers, volume 2: Mind, language, and reality*. Cambridge University Press

Putnam, H., & Goodman, N. (1983). Foreword to the Fourth Edition. *Fact, fiction, and forecast* (4th ed.). Harvard University Press.

Quine, W. V. (1969). Natural kinds. *Ontological relativity and other essays* (pp. 114–138). Columbia University Press. https://doi.org/10.7312/quin92204-006

Schrenk, M. (2016). *Metaphysics of science: A systematic and historical introduction*. Routledge. https://doi.org/10.4324/9781315639116.

Slater, M. H., & Borghini, A. (2013). Introduction: Lessons from the scientific butchery. In M. O'Rourke, M. H. Slater, & J. K. Campbell (Eds.), *Carving nature at its joints: Natural kinds in metaphysics and science* (pp. 1–32). MIT Press

Stich, S. P. (1990). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation*. The MIT.

Strößner, C. (2022). Criteria of naturalness in conceptual spaces. *Synthese*, *200*(78), 36.

Thompson, E. (1995). Colour vision, evolution, and perceptual content. *Synthese*, *104*(11 July), 1–32. https://doi.org/10.1007/BF01063672.

Part II – Publications

Paper B

Conceptual spaces: Naturalness or cognitive sparseness?

Bibliographic reference. Scholz, S., & Vosgerau, G. (2025). Conceptual spaces: Naturalness or cognitive sparseness? Synthese, 205, 129.

DOI. https://doi.org/10.1007/s11229-025-04941-8

Author contribution. Sebastian Scholz conceived the article. Gottfried Vosgerau wrote Sections 2.2, 4, and 6. Sebastian Scholz wrote all other sections. The idea underlying Section 5 and some of its wording were developed jointly by Sebastian Scholz and Gottfried Vosgerau.

Synthese (2025) 205:129
https://doi.org/10.1007/s11229-025-04941-8

ORIGINAL RESEARCH

Conceptual spaces: Naturalness or cognitive sparseness?

Sebastian Scholz[1] · **Gottfried Vosgerau**[1]

Received: 19 May 2024 / Accepted: 23 January 2025

Abstract
The conceptual spaces framework posits that conceptual content is structured geometrically, and is equipped with cognitive criteria of naturalness (namely, convexity and principles of cognitive economy). Its proponents suggest that cognitive naturalness is naturalness simpliciter, a novel move in a debate that is traditionally focused on how the world, and not the mind, is structured. We argue that "cognitive naturalness" is a misnomer and that the framework describes cognitive sparseness instead. To demonstrate this, we explore the approach's shortcomings across various branches of the naturalness debate, most notably its failure to distinguish natural kinds from fictional kinds. Our diagnosis is that the evolutionary pragmatism employed by its proponents fails to establish a connection to the real world, thus failing to secure the ontological and epistemic objectivity required for a theory of naturalness. We propose an alternative view, ecological empiricism, which posits that natural concepts or properties are those revealed through interaction with the real world.

Keywords Concepts · Naturalness · Objectivity · Ecological empiricism · Semantic Externalism · Metaphysics of Science

1 What is cognitive naturalness?

Traditionally, naturalness is *ontological*. It concerns the world having objective or mind-independent joints to be discovered, or being carved with "conceptual cutlery" (Slater & Borghini, 2013: 25). This rather gruesome metaphor harks back to Plato's

✉ Sebastian Scholz
s.scholz@hhu.de

Gottfried Vosgerau
vosgerau@hhu.de

[1] Institut für Philosophie, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Published online: 12 March 2025

Phaedrus, but the underlying idea has carried over into contemporary analytic philosophy, where the semantic externalists repopularized it in the 1970s.

Cognitive naturalness is quite different. As a first approximation, whether an entity is natural here depends on properties of the mind instead of the world. Again, the idea has been around for some time. It relates to John Locke's *nominal essences*, which "[…] are the abstract ideas that constitute the definitions of species or genera" (Jones, 2023: 1)– but in its modern form emphasizes cognitive psychology over naming and language. Here is David Lewis, a contemporary metaphysician with considerable influence on the naturalness debate, with a disparaging comment:

> Nor should it be said […] that as a contingent psychological fact we turn out to have states whose content involves some properties rather than others, and that is what makes it so that the former properties are more natural. (This would be a psychologistic theory of naturalness). (Lewis, 1983: 377).

Such psychologistic theories have a natural tendency towards *conventionalism*, which is "[…] the view natural kinds don't exist independently of the scientists and others who talk about them." (Bird & Tobin, 2024: 1.1.2). If *strong* conventionalism is true, then there simply are no joints of nature and our classifications are all relative, if not arbitrary. This squares very poorly with Lewis' traditional approach, which will be fleshed out in some detail below. The traditional theories will be called *strongly realistic*, where "realism" shall not be read as an endorsement of abstracta or universals, but of objective or mind-independent joints of nature. The account of cognitive naturalness we discuss, however, combines *weak* realism with *weak* conventionalism, i.e., naturalness is partly about the world and partly about how we think of it. It was penned by Peter Gärdenfors, the first and to date only researcher to develop cognitive naturalness *criteria*. We will therefore take a closer look at his theory of Conceptual Spaces (CS) and examine whether it is relevant for the metaphysical problems that are pertinent to the debate. Next, we will explore what the account says about various distinctions that are important to philosophers, including discussions on mind-independence and objectivity. We conclude that CS' criteria primarily reflect cognitive sparseness, and we propose an alternative view to address its shortcomings. quer

1.1 Conceptual spaces: the framework

Conceptual Spaces are theoretical entities that are postulated within cognitive science to explain and predict cognitive phenomena, or to construct artificial agents. They are mathematically defined, as they are based on axiomatic geometry, metrics, dimensions, and the like. Think of them as models of cognition which epitomize the proposition that conceptual content is structured geometrically. Crucially, these models represent the *similarity* of objects by their spatial proximity in CS– which is also how naturalness enters the scene: According to Quine (1969), similarity and naturalness are variants of the same notion. It is their similarity in relevant respects that makes members of a natural kind members of that kind. However, he did not find a satisfying analysis of either notion, which Gärdenfors ascribes to his reliance on the meth-

odology of logical positivism. By contrast, the geometric analysis seeks to "go below language" (Gärdenfors, 2003: 95) and is presented as a remedy to Quine's problems. As similar objects form clusters in CS, the degree of cohesion of the respective regions can be taken to indicate their degree of naturalness. A necessary condition of naturalness is derived, namely *convexity*, which expresses a very high degree of cohesion. A region is said to be convex, iff for any points A and B from within the region, there is no point C between A and B that does not lie in the region as well.[1]

Let's illustrate the approach with Gärdenfors' (2000) own paradigm example, the *color domain*. "Domain" means that the relevant dimensions come as a package. The color domain consists of three dimensions, namely hue, saturation, and brightness, where hue is a circular dimension accounting for complementary colors. This "color spindle" may be calculated from psychological data such as similarity judgments about visual stimuli by regression analyses such as *multidimensional scaling*. Visual properties such as *being green* are defined as regions in the domain. *Being green* is convex and a good candidate for a natural property. By contrast, the gerrymandered Goodman (1955) property of *being grue* (defined as: *"applies to all things examined before t just in case they are green but to other things just in case they are blue"*, *p. 74*) turns out to be non-convex in a visual space plus time-dimension because the corresponding region disconnects at time *t* (cf. Gärdenfors, 1990). It is evident the author devises his criterion not merely for descriptive cognitive research, but for solving philosophical problems such as Goodman's New Riddle of Induction. Accordingly, the approach has implications for a whole array of sub-disciplines that will be explored in the upcoming sections.

1.2 Cognitive semantics and pragmatism

Putnam (1975) coined the slogan of semantic externalism: *Meanings just ain't in the head!* More specifically, he opposed the idea that the extensions of natural kind terms such as "water" are determined by their intensions, which are cashed out by internal psychological states. Internalism can culminate in the notion that adherents of different scientific paradigms talk past each other, which leads to radical epistemic relativism (cf. Schrenk, 2016: 238–239). By contrast, Putnam adopted a version of strong realism in which natural kind terms are thought to refer directly by being causally linked with essences such as (presumably) H_2O, so that diverging theorists are still able to talk about the same 'stuff'.

Gärdenfors strongly disagrees with Putnam. His approach runs under the label of *cognitive semantics*, which means that the referents of natural kind terms– and all other terms for that matter– are thought to be cognitive entities. Accordingly, the term "water" is thought to refer to the concept WATER, which is a convex region

[1] Douven & Gärdenfors (2019) augment the convexity criterion with principles of cognitive economy and efficiency. Convex regions are seen as the result of an optimization procedure in which limited cognitive creatures try to strike an optimal balance between informativeness, parsimony, and other desiderata in order to survive and adapt to their environments. By appealing to optimal design, the authors move towards postulating not only necessary, but sufficient criteria of naturalness. The details of this approach are included where relevant to our argument. Where they are not, we will refer to convexity as a shorthand.

of some conceptual space.[2] At first glance, this looks like a paradigm instance of semantic internalism, but there are some caveats to point out. First, he acknowledges that language has a social dimension, which Putnam argues for by appealing to the "linguistic division of labor" between experts and laypeople. The process in which conceptual spaces are formed is conceived as a "meeting of minds" (Warglien & Gärdenfors, 2013), so meaning is treated as a communal phenomenon, not as a matter of individual psychological states. Second, internalism about linguistic meaning does not necessitate internalism about mental content: *Does Semantics Need Reality?*

> […] "not directly." Once we accept the conceptual structure of an individual as given, the semantic mapping […] can be described without any recourse to the external world. But a second part of the cognitivistic answer is "indirectly," since the conceptual structure is built up in an individual in interaction with reality. (Gärdenfors, 1999: 14).

Some affinity to realism is clearly present. But does that mean our natural concepts get their meanings from latching onto essences? Absolutely not.

Gärdenfors does not want to subscribe to any view in which the concepts correspond to or represent the world, which would be typical of essentialism or other forms of strong realism. But since both conventionalism and internalism invite relativism, he does want the world to play *some* role in shaping our concepts.[3] This balancing act is undertaken by an appeal to evolution: "Via successful and less successful interactions with the world, the conceptual structure of an individual will adapt to the structure of reality." (Gärdenfors, 2000: 156). It is important to note that he does not talk about reality in and of itself here, but reality as opposing and satisfying our evolutionary needs. This, however, is a pragmatist account akin to Quine's. To put it in a nutshell: Our concepts do not refer, but they *work* because we would have died out if they did not. The 'primordial' CS are thought to provide the basis for more sophisticated, or scientific endeavors to operate on.

2 Shortcomings

We saw how the CS approach invites criticism for leading to problematic relativism, as it makes naturalness dependent on the mind. To address this concern, its proponents adopt a weak version of conventionalism (there is some leeway for conven-

[2] More precisely, it is a "[…] set of convex regions across multiple domains together with a prominence assignment to the domains and information about how the regions in different domains are correlated." (Gärdenfors, 2008: 310).

[3] To get a clear sense of how relativism affects the account, consider again how the CS come about. They are not divinely ordained but rather chosen by scientists who map and model cognition from data on similarity judgments, for instance. This begs the question: Can't we just "pick [our] own conceptual space and in this way make [our] favorite properties come out natural" (Gärdenfors, 1990, p. 91)? If even contrived concepts such as CAT-OR-ELECTRON could be rendered natural, this would lead into full-blown conventionalism– which Gärdenfors seeks to avoid. For an in-depth discussion of contrived (or "gerrymandered") concepts, see Sect. 3.

tions, but we cannot choose the conceptual division of the world at will), while at the same espousing a weak version of realism (we have no epistemic access to the world in itself, we only deal with it successfully). Natural concepts are said to be mind-dependent but still objective. Thereby, proponents of the CS approach suggest cognitive naturalness is naturalness *simpliciter*. By contrast, we will argue that cognitive naturalness is no naturalness at all.

The fourth section will take a closer look at the concepts of mind-independence and objectivity. But first we will show that cognitive naturalness cannot solve the metaphysical problems central to the philosophical debate. Crucially, it cannot account for various distinctions that are important to philosophers. The shortcomings of the approach manifest differently in different branches of the debate, so each will be considered in turn.

According to Judith Crane, "[p]hilosophical treatments of natural kinds are embedded in two distinct projects. […] The kinds studied in the philosophy of science approach are projectible categories that can ground inductive inferences and scientific explanation. The kinds studied in the philosophy of language approach are the referential objects of a special linguistic category—natural kind terms—thought to refer directly." (Crane, 2021: 12177). Her distinction will be taken up and supplemented by a third branch, which is strangely absent from her discussion: The kinds studied in the metaphysics approach are primitives that are postulated to address problems with laws of nature and causality. Note, however, that we introduce the distinction for heuristic reasons, and do not claim it reflects a deep systematic divide. The problems of naturalness can hardly be discussed in isolation.

2.1 Philosophy of language

The philosophy of language approach was shaped by semantic externalism, and is therefore closely tied to the aforementioned concept of direct reference. Proper names are thought to refer directly to the individuals they denote. Analogously, natural kind terms are taken to work as proper names for natural kinds. Thus, their meaning is not a matter of descriptive content, but of their extensions, or the external 'stuff' they rigidly designate across worlds. As with Putnam (1975) and Kripke (1980), reference is secured by causal contact with samples. Accordingly, the theories within this branch deal with the nature of the samples: What must they be like in order for direct reference to work?

Crane (2021) discusses three contemporary approaches, all of which indulge in some form of essentialism.[4] This is because essences are required for rigid designation: The essence of a natural kind is what fundamentally defines it– independently of any particular properties that may be observable in any given world. Natural kind terms would fail to pick out the same stuff across worlds, if they would latch onto the surface properties that are varied. A good deal of her discussion is about the *qua problem* of reference fixing, which has been characterized as follows:

[4] The three forms of essentialism she discusses are represented by Salmon (1981), LaPorte (2003), and Cook (1980), respectively.

> When a speaker grounds a term through perceptual experience, in virtue of what is the term grounded in the cause of that experience qua-one-kind and not qua-another? For that cause will always be an instance of many kinds; e.g. the one object may be an echidna, a monotreme, a mammal, a vertebrate, and so on. (Devitt & Sterelny, 1999: 311).

However, if "any object can belong to at most one natural kind", i.e., if *uniqueness* holds, then "natural kinds […] provide a unique partitioning of the world, carving nature at a unique set of joints" (Crane, 2021: 12191). As a result, the qua problem does not arise because there is only one right kind to ground a given term. Unfortunately, such a view is difficult to maintain because it "severely limits the kinds that count as natural, and it surely discounts the vast majority of scientific kinds" (ibid.).

You would expect the CS account to be silent on all these issues because of its deep-rooted anti-essentialism. Douven goes out of his way to dismantle essentialist intuitions, e.g. by problematizing the micro-essentialist sentiment that water is H_2O (cf. Douven, 2022: 11), and Gärdenfors suggests psychological essentialism is bad metaphysics (cf. 2000: 4.2.2). Moreover, Gärdenfors (1999) criticizes traditional semantics such as direct reference theories for not explaining how cognitive agents are able to "grasp" meanings. For cognitive semantics, this is a matter of establishing neural connections between linguistic expressions ("green") and cognitive structures (GREEN-region), where the latter are at least partly perceptually grounded. Since this view is committed to the idea that "language represents a conceptual structure, but it does not directly represent the world" (Gärdenfors, 2000: 154)– which obviously precludes direct reference theories– cognitive naturalness is simply irrelevant to the metaphysical problems discussed in the philosophy of language branch of the naturalness debate.

Interestingly enough, however, both authors are concerned with uniqueness. For instance: "In standard realist thinking, there could never be more than one conceptual scheme capturing the natural kinds. And it is not clear that the optimal design account guarantees satisfaction of this uniqueness condition." (Douven, 2022: 10). There are a couple of things to unravel here. First, note that this depiction of realist thinking is a straw man. As Crane points out, few philosophers accept uniqueness, and subsequent sections will demonstrate most contemporary realist thinkers accept some form of pluralism about organizing the natural world. Second, it is not obvious to us why the account would need uniqueness in the first place. The reason cannot be to address the qua problem, as this issue arises specifically within causal theories of direct reference, which the CS account does not employ.[5] Most probably, however, Crane and Douven speak of two different kinds of uniqueness: While Crane talks about the idea that for each object there is only one natural kind which it belongs to, Douven addresses the relativistic worry that there might be more than one conceptual scheme. If spaces can be partitioned in multiple optimal ways, e.g. by rotating

[5] We may still ask how the right conceptual region is assigned to a term like "echidna", which would be the cognitive analogue of the qua problem. Presumably, the answer is that linguistic expressions are annexed to cognitive entities in a more dynamic way than to essences. Reference might even change depending on the context, etc.– one does quite certainly not need uniqueness to solve this one.

an optimally partitioned space, then naturalness may look like an arbitrary business, independently of how many kinds are assigned to each object. For this reason, Douven (2022) claims that the optimal partitions are just a very small, *sparse* subset of the possible partitions. However, this does not relate to the discussion about naturalness in the philosophy of language.

2.2 Metaphysics

The discussion about naturalness in metaphysics is mainly inspired by the philosophy of Lewis (1983, 1986), who introduces the notion to solve several interrelated problems. Some of these problems are rather problems within the philosophy of language. We concentrate here on the metaphysical problems (which are, of course, not unrelated to language).

In order to formulate a theory of laws, Lewis introduces his "best system analysis (BSA)". Based on the doctrine of Humean Supervenience (Lewis, 1986), his idea is roughly that laws are the axioms of those deductive systems that best describe the Humean mosaic. The trouble, however, is that there might be an abundance of such deductive systems that turn out equally well but differ in their vocabulary (Loewer, 1996). Moreover, not restricting the possible vocabulary might also lead to the technical difficulty resulting in one single law (Loewer, 2007).

In order to give a metaphysical account of causation, Lewis (2000) presents a counterfactual analysis of causal influence. Combined with his modal realism (Lewis, 1983), this leads to the problem that for every object a proper counterpart or duplicate in other possible words must be found.

Both problems can be solved with "perfectly natural properties" (Lewis, 1983): they determine the vocabulary to be used in the best system and they determine duplicates, which are those objects that share (almost) all perfectly natural properties. As we saw in the quote at the beginning of the paper, Lewis himself thinks that any psychologistic account of natural properties would be a non-starter. However, proponents of the "better best systems account" (BBSA) like Schrenk (2017) and Cohen and Callender (2010) have argued that at least the technical difficulty could be solved with any random set of predicates, including "psychologistic" ones (they advocate to take basic predicates of the special sciences as starting points). Nevertheless, the rest of the job– namely to refer to the right properties– is done by the world. The idea here is that if the predicates are not referring to the right properties, the according system will never be a good one, let alone the best (or a best, according to BBSA).

This kind of thought seems to be what Gärdenfors (2000) has in mind when he says that our concepts are based on our interaction with the world. Douven (2024) explicitly relates CS to Lewis's BSA, suggesting that CS could provide a solution to the problems described above. Nevertheless, we are not convinced that this could work for a definition of natural properties. The reason is, as we have argued, that the world itself must play a pivotal role in the determination of which predicates (or concepts for that matter) can be called natural.[6] Thus, any account that purely concentrates on

[6] As Sider (2011: 27) puts it: "Lewis's constraint on reference is "externalist"; reference is not determined merely by us."

cognitive criteria could maybe establish suitable systems of predicates. However, in order to solve the metaphysical problems, we additionally need the requirement that they refer to the right kind of properties, and this can never be determined by cognitive criteria alone. Merely gesturing at some evolutionary successful interaction is not enough, because this pragmatic move is far too unconstrained.[7]

2.3 Philosophy of science

This branch of the naturalness debate revolves specifically around grounding induction and scientific explanation, as envisaged by Richard Boyd.[8] Boyd's pivotal insight is that the natural world is somewhat messier than the essentialists would have us believe. Traditional essentialism assumes that kinds are defined by necessary and jointly sufficient conditions. By contrast, modern evolutionary biology recognizes that species and taxa are not immutable, uniformly characterized groups. Instead, they are dynamic entities, subject to the forces of genetic drift, natural selection, and gene flow, which can lead to the emergence of new traits and the loss of others without necessarily leading to the formation of a new species. Thus, Boyd conceives of kinds as *homeostatic property clusters*. The idea is roughly that some properties like to 'stick' together qua causal structure to form stable patterns and ground our scientific endeavors. These patterns are objective and make the kinds what they are, while their historicity and vagueness are fully acknowledged.

We are not going to discuss the strengths and weaknesses of Boyd's and his successor's accounts here. What is remarkable, however, is that his thinking, as well as the broader debate he initiated, have been undergoing a pragmatic turn in the sense that the role of scientific practice has been increasingly strengthened, while realism has been losing significance. The early Boyd was basically an essentialist and a direct reference theorist, while the later Boyd relativizes naturalness to *disciplinary matrices* and embraces pluralism in the sense that there are multiple legitimate ways to carve up nature. Although Crane lists realism as one of the key characteristics of the philosophy of science accounts, the realism some of them employ is quite minimal:

> Scientific realism requires that true scientific theories, and the concepts and categories they employ, are responsive to real aspects of the world. […] The more pragmatic accounts require only that a scientific kind be useful for scientific practices, while less pragmatic accounts require that they be connected to causal structures or mechanisms. (Crane, 2021: 12194).

You would think this is where the CS account fits in and can finally shine. Gärdenfors is clearly concerned with grounding induction in naturalness– recall Goodman's Riddle. The account resonates with other theories that make epistemic subjects part of the naturalness equation, e.g. Boyd's disciplinary matrices, Chang's (2022) opera-

[7] In contrast, the requirement of success of a scientific theory might fare much better since scientific methodology is constrained in a way that aims to secure a truth-conducive relation to reality. See also Sect. 4.

[8] "It is a truism that the philosophical theory of *natural* kinds is about how classificatory schemes come to contribute to the epistemic reliability of inductive and explanatory practices." (Boyd, 1999: 146).

tional coherence, or Massimi's (2022) perspectival realism. It might even have an edge over its competitors because it uniquely comes with empirically testable naturalness criteria. Yes, its notion of naturalness is purely cognitive and contains descriptivist elements, but due to its pragmatism is still "responsive to real aspects of the world"– or is it?

We see at least two reasons to doubt that the account can offer meaningful solutions to the problems of the philosophy of science branch of the naturalness debate: Its reliance on evolution and its focus on perception. The first point relates to Goodman's Riddle, which is one of the central problems of the debate. Scholz (2024) argues that Gärdenfors' solution strategy amounts to grounding the convex concepts' instrumental success in their evolutionary success, thereby replacing one relativism with another. Even if this move can be spelled out in a non-circular way, it invites evolutionarily motivated epistemic skepticism because *poor reasoning can be highly adaptive*. In other words, there is an insurmountable gap between what is evolutionarily useful and what the world is really like– just think of the contrast between the color concepts and the physical light spectrum. This example is also the transition to our second point, namely that the CS account of naturalness works well only for cases that are close to perception. Its standard examples involve color, taste, sound, etc., but the further removed from perceptual domains the concepts are, the more difficult it is to find interpretable dimensions and plausible naturalness criteria. Scientific concepts, however, are frequently very abstract, hard to learn, and cannot easily be traced to perceptual states– as also acknowledged by Douven (2024: 8–9). Thus, Boyd and his successors would likely regard the CS account as irrelevant to their theoretical ends. Note that Gärdenfors and Zenker (2013, 2014, 2016) have published a number of papers on how scientific theory structure and change can be represented in the CS framework, but these papers do not mention naturalness criteria at all– which is quite revealing from our point of view.

The deeper issue underlying both points is that the account is not only responsive to real aspects of the world, after all. It is responsive to how we evolved to perceive and think about the world. The world as described by physics, for instance, is quite different from what everyday experience tells us. Science can help us overcome our evolutionary misconceptions and biases precisely because it picks up on what the world is really like.[9]

3 Kinds of kinds

Over two thousand years of naturalness debate have brought about a whole zoo of putative kinds of kinds. Philosophers have typically introduced them as contrasting classes to those considered as natural, serving diverse theoretical purposes. This section is devoted to exploring some of these distinctions in order to review whether and how the CS criteria trace the ontological differences that are at stake. To spoil

[9] Of course, our perception is also able to track some causal relations in the world, but they will always depend on the causal mechanisms of perception and will thus be potentially mind-dependent in ways that make them unsuitable for the needs of the Philosophy of Science project.

the discussion a bit, it turns out that the account can only really trace one distinction, namely the one between sparse and gerrymandered kinds. For most cases, this is not necessarily a problem, but we do think there is one highly problematic case, namely fictional kinds.

Gerrymandered kinds are exemplified by grue emeralds, non-ravens, cats-or-dogs, fungi-that-grow-within-a-100 m-radius-of-my-house, and so forth. They figure prominently in the philosophy of science debate because in contrast to the natural kinds they are thought *not* to support scientific reasoning and explanation. For instance, non-black non-ravens are generally taken *not* to confirm the hypothesis that all ravens are black, although, technically speaking, they do (cf. Hempel, 1945). The trouble is this: When we permit contrived predicates or concepts that denote gerrymandered kinds and properties to be projected in reasoning, we face a scenario where anything can seemingly confirm anything else. Consequently, the standard instantial model of confirmation, according to which a "positive instance of a generalization lends some support to that generalization" (Slater & Borghini, 2013: 5), breaks down. (This is the main reason why Goodman's Riddle is such a big deal). Ruling out the undesirable predicates, however, is not straightforward. Many philosophers think the good predicates are precisely those that carve nature at its joints, while the bad predicates refer to gerrymandered kinds, if anything at all. Within the metaphysics branch of the naturalness debate, the gerrymandered kinds are a thorn in the side of *set nominalism*. Set nominalists such as David Lewis take any set, or arbitrary combination of objects to count as a full-blown property or kind. This "abundance" of properties and kinds leads to the theoretical problems discussed above. Thus, Lewis delineates among all the groupings of things in nature an elite, or *sparse* subset of natural groupings of similar things to serve his theoretical ends.

Gärdenfors seems to achieve a similar result by different means. As was previously mentioned, convexity successfully precludes the Goodman property of *being grue*. Other cases are tricky, e.g. the property of *being non-black* is presumably a convex region of the color domain. (This is not detrimental, of course, as convexity is meant to be necessary, but not sufficient.) Whether the relevant examples are represented by non-convex regions is an empirical concern. However, we consider it plausible that the gerrymandered kinds can be eliminated by *some* criteria of cognitive economy. Why is that? Precisely because they do not depend on the structure of reality (or do so only in a derivative sense when some of their components do), but are artifacts of our own mental "contortions". For instance, we can always conjure up contrived concepts by performing logical operations, e.g. by combining natural ones into random alternations. It goes without saying that such concepts strike a poor balance between informativeness and parsimony, and are difficult to process. Thus, it should come as no surprise that the CS account is well-suited for dealing with these cases. The fact remains, however, that this makes the ontological difference between sparse and gerrymandered concepts a matter of them being represented in different ways. Although feasible, realistically minded philosophers will not be content with this way of drawing the distinction.

Functional kinds are important especially to the essentialists of the philosophy of language branch because their nature is a matter of function, rather than any deep underlying features. One and the same function can be fulfilled by vastly differing

properties, which is to say functions are multiply realizable. Thus, they are puzzling for direct reference theorists because it is unclear what the functional kind terms latch onto. For instance, Schwartz (1978)– who is otherwise an adherent of semantic externalism– denies that artifact terms such as "pen" refer directly. Instead, he speaks of *"*nominal kinds*"* to associate with Locke's nominal essences.

Gärdenfors (2000) acknowledges that "the analysis of functional properties is an enigma for [his] theory" (p. 98) because CS are all about perceptual properties. He briefly explores reducing them to "the actions that the objects 'afford'" (ibid.), and extends this analysis in a later paper (2007). Independently of whether this is feasible, it should be clear that his naturalness criteria are not able to differentiate between functional and other concepts, as natural action concepts are thought to be represented in terms of convex regions as well. On the other hand, one might argue that the account can mimic essential features by placing maximal salience weights on the respective dimensions. However, such maximal salience weights would equally apply to natural kinds (e.g. water having a certain chemical structure) and to functional or "nominal" kinds (e.g. chairs being something to sit on). Since the result of maximal salience weights are not essences but features which are *represented* as being essential, this move cannot provide the ontological distinction in question.

What is so striking about artifacts is that they are *made by us* for specific purposes. This circumstance puts them into dubious ontological standing– from the point of view of traditional realist metaphysics, that is (cf. Khalidi, 2016). Traditional realists tend to stress the importance of mind-independence when it comes to judging how real something is. The next section will discuss this notion in some detail. For the time being, note that many kinds of kinds are mind-dependent in some sense, and that, consequently, naturalness criteria may be evaluated on whether they delineate them from the natural ones. This concerns, for example, categories that are known from social ontology, namely *social kinds* such as money or democracy. Another potentially problematic class are *psychological kinds* such as belief, anger, depression, and so forth. Both might be considered functional alongside artifacts, but there are non-functional cases as well, e.g. *fictional kinds* such as orcs or gods. Moreover, there are a number of intriguing borderline cases, most notably *artificial kinds* such as dog breeds or synthetic chemicals, that appear to be mind-dependent in some sense of that term, but might be taken to have essential features.

How does the CS account deal with these kinds of kinds? Douven (2024) does not make an ontological distinction between natural and social concepts, which is indicative of a larger issue: Although the details are open to empirical inquiry, all of the examples given above can *prima facie* be expected to be represented by cognitively economic concepts, so likely all of them will turn out CS-natural. While problematic from the point of view of traditional realism, this is not necessarily detrimental. One might argue that social, psychological, and artificial kinds are natural phenomena open to scientific inquiry, so no "hard" ontological distinction is required. But this is not an option in the case of fictional kinds: Entities like orcs, wizards, and gods in myths, which do not have a physical existence. They exist primarily in narratives and are products of human imagination and do not causally interact with the physical world. Knowledge about these kinds is derived from narrative contexts and artistic creations. They are not bound by empirical reality but by the internal logic and rules

set by their creators or the traditions of their genres.[10] Qua definition, they have nothing to do with the joints of nature, so any theory of naturalness that treats them as natural must be regarded as highly revisionary. Note that the case of fictional kinds constitutes a great illustration for a point made in the preceding section: Fictional kinds can only turn out natural in an account that is not– or at least not only– "responsive to real aspects of the world".

One may wonder, however, whether Douven's adaptation of the Best Systems account can come to the rescue. After all, part of the motivation behind his move is to make scientific concepts such as those of the social sciences accessible to the CS criteria.[11] One could perhaps argue that fictional kinds– as opposed to social kinds– do not participate in any best system and are therefore not objective, which allows demarcation. However, we will argue that CS-natural kinds are not objective either.

4 Mind-independence and objectivity

Douven (2024) argues that while natural concepts as defined by conceptual spaces "are mind-dependent in *some* sense of that expression, mind-dependence in this sense does not compromise their objectivity" (p. 13). Traditionally, objectivity is closely related to mind-independence in the following sense: Something is said to be objective(ly real) if it is as it is independently of how we conceive of it, i.e. if it is so independently of human minds. When applied to concepts, the idea behind the criterion is that objective concepts "carve up nature at its joints", such that nature itself dictates the boundaries of the concepts. However, both the notion of mind-independence and the notion of objectivity are rather vague and have to be spelled out in more detail to evaluate this claim.

Let us begin by distinguishing two notions of mind-independence which are frequently mixed up– to the detriment of the debate. Firstly, things could be said to be independent of minds if their existence is independent of the existence of (human) minds. This will quite obviously not do for natural concepts: if TIGER and DOG are natural concepts (these are standard examples), so is HOMO SAPIENS. However, the existence of homo sapiens is not independent of the existence of human minds, since humans actually have a human mind (for a similar argument, see Khalidi, 2016).

Much more plausible is to say that, secondly, something is mind-independent if it is as it is independently of how we cognize (perceive, think about, linguistically describe) it. Homo sapiens are the way they are, no matter what theory we have about them. The same is even true for perception and cognitive processes (thinking): Pre-

[10] This is not to say that fictional kinds never relate to *any* aspects of the physical world. However, such relations will not be central to the kinds in the sense that fictional kinds cannot be explored by investigating the world.

[11] "This is proof of principle that we can construct conceptual spaces for scientific concepts– in the above case, concepts pertaining to the social sciences– in basically the same way in which we can construct a space for color concepts […]. […] In short, the question of how we are to demarcate bona fide scientific concepts from gerrymandered ones is not readily provided by Douven and Gärdenfors' optimal design principles. The following proposes an answer to this question that retains the spirit of Gärdenfors' optimality approach by adapting Lewis' Best Systems Account." (Douven, 2024: 8–9).

sumably, humans perceive and think the way they do independently of which theory of perception or thought is currently available.[12]

To see how these ideas connect to the idea of objectivity, we follow the helpful distinction by Sankey (2024) between ontological objectivity, objectivity of truth and epistemic objectivity, which he introduces for his discussion of the objectivity of science. Only the first two notions relate to mind-independence, the third one is concerned with epistemically fruitful methods. Moreover, since we are concerned with concepts here which cannot be true or false, the objectivity of truth is not very important for our discussion and is thus only mentioned alongside with ontological objectivity.

4.1 Ontological objectivity

The notion of an "'objective reality' [expresses] the idea that the world or reality exists in and of itself […] independently of human belief, thought or experience." (Sankey, 2024: 2) Applied to concepts this means that a concept picks out parts of objective reality if what it refers to (its extension) is what it is independently of how we cognize it. TIGER and HUMAN MIND are such objective concepts, since tigers (the kinds or the individuals) and human minds (the kinds or the individuals) are the way they are independently of our cognition about them.[13] MONEY and DEMOCRACY, for example, are not objective because something is only money or a democracy if people have certain beliefs about them.[14]

To say that there is "objective truth" is to say that "[t]ruth is objective in the sense that it does not depend on what we believe." (Sankey, 2024: 4) The connection between ontological objectivity and objectivity of truth is easy if we accept some kind of correspondence theory of truth: If a sentence/thought/proposition is objectively true, it is so because the world is as the sentence/thought/proposition states independently of how (or if) we think about it. This requires that the concepts occurring in the thought pick out parts of objective reality.[15]

Concepts are ontologically objective only in a derivative way, namely if they refer to objective kinds. Thus, the "work" to make them objective is entirely done by the world, the concepts have no role to play in it except referring to the right thing. But how they exactly refer is of no importance. It could be because of some deictic

[12] Note that, unless you have platonic intuitions about concepts, all concepts are at least highly mind-dependent (if not mental themselves). Thus, natural *concepts* could never be mind-independent; rather, concepts can only be said to be mind-independent in a derivative way if they refer to mind-independent kinds.

[13] What exactly the extension of a concept is, is left open on purpose: Some might want to say that natural kind concepts refer to natural kinds which have their members or their essences independently of our cognizing about them. Others may prefer a more nominalist perspective and say that they refer to classes of things which have their properties independently of us.

[14] Of course, single individuals could be wrong about whether they live in a democracy or not. However, nothing can be a democracy without some people having beliefs about free voting, for example. In this sense, unless at least some people have certain beliefs, there will be no democracy.

[15] Things might be different with other theories of truth (for a discussion, see Sankey, 2024). However, since the idea behind natural concepts is that they refer to something in objective reality, the very idea of natural kinds and natural concepts strongly favors a correspondence theory of truth.

introduction of the term– a baptism in the sense of Kripke (1980)– because of some descriptive content or whatsoever. None of these mechanisms, however, is able to determine whether the concept is ontologically objective or not. Take for example color concepts, that are most likely introduced by deixis. As Putnam (1987: 4–8) argues, it does not follow that they track objective properties; rather, they track complex properties that are dependent on how we perceive surfaces.[16]

Therefore, a theory of concepts will never be able to determine whether concepts are ontologically objective (in the derivative way). Only a theory of the objective kinds that the concepts refer to could possibly shed light on the status of concepts regarding ontological objectiveness. Moreover, mind-independence does play a crucial role for ontological objectiveness. Thus, Douven (2024) must have a different kind of objectivity in mind when saying that the theory of conceptual spaces renders concepts objective though mind-dependent in some sense.

4.2 Epistemic objectivity

Epistemic objectivity is related to methods in the case of science, according to Sankey (2024). Translated to the case of concepts, epistemic objectivity is related to the processes of concept formation. Douven (2024) is likely to have this kind of objectivity in mind, since his view is "close to Putnam's (1981) internal realism" (p. 4), which he claims to be based on the idea that our conceptual apparatus is made up in a way that ensures objectivity. His characterization of conceptual spaces, he claims, is such that they are governed by rules that are not under our control. Thus, "to say that natural concepts are mind-dependent in the sense of Douven and Gärdenfors (2019) does not imply that it is dependent on our thinking and theorizing what the natural concepts are." (p. 5) This point is very much in line with our argument above. However, the critical point is whether this form of mind-independence ensures epistemic objectivity.

Sankey (2024) discusses two problems of epistemic objectivity, namely the theory-dependence of observation and the variability of the methods of science. He concludes that science is epistemically objective because the methods are "highly reliable truth-conducive tools of inquiry that the sciences lead so regularly to successful interactions with the world" (p. 9). The important point for our discussion here is that epistemic objectivity can only be obtained if a certain connection to the real world is guaranteed by the epistemic tools we use. A different point to the same effect is made by Khalidi (2016: 243): "In distinguishing real from bogus kinds, what concerns us is whether a kind exists in the sense that it has instances that share causal properties." Although he makes no claim about the methods used to investigate the causal properties, his criterion of their being real instances of the kind plays the role of ensuring a stable connection to the actual world.

Applied to theories of concepts, epistemic objectivity can thus be established if the cognitive processes of concept formation ensure a stable connection to the real

[16] To be sure, this is not to say that color concepts are not related to *any* causal structure of the world; it is only to say that they are also dependent on how our perception (i.e. mind) works and are thus not what they are independently of how we cognize them.

world. In other words, the formation processes must be such that they are guaranteed to track the right kinds in the real world. Douven (2024) might be right that his theory ensures one optimal and sparse conceptual scheme; however, since all his criteria for cognitive naturalness are internal, they cannot establish the required connection to the world. Again, the case of color concepts is a formidable example. In short, criteria of concept formation can never be epistemically objective if they stay internal and do not include some mechanism to establish the connection to the real world.

The appeal to evolutionary success will not suffice, since evolutionary success is much too unconstrained to guarantee the tracking of objective reality. As argued for above, evolution is only sensitive to those aspects of the world that are connected to our needs. Thus, we are not tuned to reality "as it is by itself", but to reality as it is supporting or hindering our survival. Epistemic objectivity, however, asks for more: It requires truth-conduciveness, not survival. What CS is lacking is thus cognitive mechanisms behind concept formation that are truth-conducive, i.e. that are able to track causal relations in the world which are as independent of our cognitive apparatus as possible. In the conclusion, we will shortly hint at Ecological Empiricism as a theory to describe such cognitive mechanisms.

5 The problem of abundance and the problem of objectivity

We have argued that the CS account is not able to provide criteria for naturalness, at least not in a way to satisfy the philosophical needs naturalness was designed to address. However, we do not wish to argue its criteria are thus worthless– quite on the contrary, CS is the first and only theory to spell out cognitively adequate criteria of sparseness.

To evaluate this achievement, we must first take a few steps back. The concept of sparseness has been mentioned above in connection with Lewis' set nominalism, namely as a solution to the problem of abundant properties. However, the problem is not specific to set nominalism, but is discussed more broadly. Hempel encounters a version of the problem, as the received theory of confirmation breaks down if abundant predicates such as "… is a non-raven" are allowed into scientific reasoning. It also concerns Armstrong (1978), who "used the traditional doctrine of universals to draw the distinction between genuine and nongenuine features […]: there simply is no universal of 'being either a cow or an electron'." (Sider, 2011: 4). Both Lewis and Armstrong address the issue by invoking naturalness: "Call a predicate "sparse" when it marks a joint in nature. For Armstrong, a predicate is sparse when there exists a corresponding universal; for Lewis, a predicate is sparse when there exists a corresponding natural property or relation." (Sider, 2011: 85). For Gärdenfors, however, a predicate is sparse when it picks out a convex concept.[17] As we have argued, this notion of sparseness cannot be about the objective joints of nature. Rather, we should understand Gärdenfors as demonstrating that there are other means to get a sparse subset of the abundant predicates than by invoking naturalness. Thus, "cognitive

[17] As mentioned above, convexity is not the only criterion (see also Douven and Gärdenfors, 2019).

naturalness" is a misnomer for the criteria he champions. To avoid terminological confusion, we should stick to "cognitive sparseness" instead.

This result indicates that two problems, which are often treated in one go, must be distinguished: The *problem of abundance*, and the *problem of objectivity*. Technically speaking, any criteria that delineate an elite subset among the abundant groupings of things will count as a solution to the problem of abundance. (This is one of the ideas behind the Better Best System Account; Callender & Cohen, 2010; Schrenk, 2017). What is great about empirically corroborated cognitive criteria, however, is that they reflect how cognitive agents actually think and reason. Consequently, such criteria are helpful in predicting the behavior of cognitive systems, or mimicking their behavior with artificial systems. Crucially, this works independently of 'what holds the world together at its core'.

This is not the case for solutions to the problem of objectivity. This problem shows up in theories of direct reference (Sect. 2.1), Goodman's new riddle of induction (1955), Lewis's (1983) need for defining counterparts of objects in possible worlds (Sect. 2.2), as well as in the philosophy of science branch of the naturalness debate (Sect. 2.3). What all these cases have in common is: They need to make sure that certain predicates, and especially natural kind terms, pick out the– or at least some[18]– *right* objective properties, thereby tracking essences, allowing for induction and establishing real similarity relations. Without addressing the problem of objectivity, essentialists are unable to fix direct reference across worlds and to address the qua problem. Realist metaphysicians are incapable of generating robust accounts of similarity, laws, causation, and so forth. Philosophers of science are powerless to ground or justify induction and scientific explanation.

To see this more clearly, let's zoom in on the latter of these points.[19] Goodman's Riddle, which is pertinent here, cannot be resolved by appealing to the simplicity of the predicates because simplicity is language relative: From within a language that contains "… is grue" as primitive vocabulary, "… is green" appears more contrived. Goodman himself embraced the notion that projectibility is thus relative, and that we just happen to project "… is green" because it got "entrenched" (cf. Cohnitz & Rossberg, 2024: 5.4). Thus, he did not think a solution to the problem of objectivity was available, in the sense that there are no objectively right predicates referring to the objectively right properties, and he left scientific reasoning ungrounded. But he did indeed delineate a sparse subset among the abundant predicates, namely those that are well-entrenched, thereby addressing the abundance problem. This is in stark contrast to the realists of the philosophy of science branch, who postulate objective joints of nature for scientific reasoning to take hold on, thereby establishing a suitable connection to the real world. For these theorists, there are objectively right predicates, namely those that refer to the right kinds or properties. Through this move, which is

[18] We do not want to preclude the possibility of multiple objectively right ways to carve up nature (e.g. Boyd's disciplinary pluralism).

[19] For reasons of length, we cannot go into the details of all these points. Just to gesture at another one: Even if we have sparse grue-predicates, we cannot establish the right similarities (within and across worlds), as the set of grue things contains members that are not similar to each other in any relevant way, which is relevant to questions of laws and causation.

quite common in the naturalness debate, they address both the problem of abundance and the problem of objectivity at once by invoking naturalness.

Scholz's (2024) paper and our discussion show that Douven and Gärdenfors want to follow this very blueprint, only without buying into traditional strong realism. In order for that to work, however, they must somehow establish a suitable connection to the real world– which they try to accomplish by deploying an evolutionary pragmatism. We argued that their approach fails to accomplish the desired goal. However, we do not want to deny the feasibility of pragmatist accounts per se. We see two general ways to ensure the right connection to the world: The first way is to let the world itself take the lead, as it is done e.g. by causal theories of reference à la Kripke (1980), reference magnetism accounts (as adopted by Sider, 2011: 3.2) or by postulating naturalness as an ontological primitive (Lewis, 1983). This way leads to ontological objectivity and objectivity of truth (Sect. 4.1). The second way is to spell out constraints on our epistemic access to the world that ensure that this access is truth-conducive, which leads to epistemic objectivity (Sect. 4.2). The latter way is taken– at least implicitly– by accounts that propose to let science take the lead in identifying the right predicates, as scientific methods are widely taken to be the prime candidate of truth-conducive methods. Thus, the second view is compatible with scientific pragmatism. What we criticize is that the CS brand of evolutionary pragmatism is too unconstrained to generate truth-conducive methods, and solve the problem of objectivity. This is the deeper reason why the CS account does not apply to the metaphysical problems of the naturalness debate. This point, as well as the distinction between the problem of abundance and the problem of objectivity, can be understood as a supplement and a substantive extension of Scholz's criticism.

6 Conclusion

Discussing the core problems that are traditionally discussed in relation to the naturalness of kinds or predicates, we argued that the theory of Conceptual Spaces (CS) is not able to offer suitable solutions. The reason is that CS only addresses one of two problems that have been tried to be solved with naturalness. The first problem is the *problem of the abundance* of predicates and kinds, which shows up in different flavors for different approaches, e.g. for confirmation theories (Hempel, 1945), for a theory of universals (Armstrong, 1978) and for Lewis's (1986) Best Systems Account. This problem is best solved by introducing "sparse" predicates, and it is not very important (for this problem), which predicates exactly are chosen.

In fact, CS proposes criteria that lead to a set of sparse predicates. Moreover, it has the advantage to spell out clear, empirically corroborated criteria and thus to provide a tool to define sparse predicates that are directly related to human cognition. We are convinced that this is an advantage over other proposals, since cognitively adequate criteria are useful for facilitating the explanatory, predictive, and constructive aims of cognitive science. Thus, we argued, CS is a very promising account of sparseness.

However, sparseness provides no solution to the second problem, namely the *problem of objectivity*. The common underlying difficulty here is that we have to ensure that our predicates pick out the right objective properties. This problem has

to be solved by securing the right connection to the world. And exactly this cannot be done by a cognitive theory like CS, since such theories simply do not speak about the world. The rather vague appeal to evolutionary success is far too unconstrained to secure a stable connection to the world.

Naturalness is indeed a solution to both problems at the same time. However, it comes with its own difficulties that have been discussed in the literature. Sparseness, on the other hand, is only a solution to the first problem, not to the second, and therefore cannot replace or substitute naturalness. What is needed for a new conception of naturalness is an approach that provides sparseness *and* the right connection to the world at the same time. In this sense, CS can be viewed as providing one part of such a new conception, but it needs to be supplemented.

There are two ways to ensure the right connection to the world: The first way is to let the world itself take the lead, which leads to ontological objectivity (Sect. 4.1). The second way is to spell out constraints on our epistemic access to the world that ensure that this access is truth-conducive, which leads to epistemic objectivity (Sect. 4.2). Since the second way sits much better with accounts that are somewhat pragmatist in spirit, as CS utterly is, we think that the best supplement for CS will take the second route as well.

As an outlook, we will sketch why Ecological Empiricism (EE; Vosgerau, 2024) is a suitable supplement for CS to yield a full account of naturalness. The basic idea is that concept formation is grounded in the interaction with the world, which provides a stable connection with the real world. In this account, natural concepts are those that refer to "properties that are revealed to us by interacting with the [real] world" (Vosgerau, 2024: 3). In the spirit of (a naturalized version of) affordance theory, EE proposes that our concepts are based on learning stable correlations between our actions and our sense input. Such correlations typically stem from causal features of the world *and* the causal make-up of our cognitive (and corporeal) apparatus. Some correlations, however, are highly dependent on our actions and are thus relatively weak, especially "social affordances" based on conventions (for example that spaghetti are eaten with fork and spoon). Other correlations are much more stable since they depend mainly on the world itself, for example that a rod of a certain length cannot be carried through openings narrower than the rod's length. Vosgerau (2024) argues that scientific measurement is a special and very sophisticated way of interacting with the world that is designed to reveal the most stable correlations. The methods applied in measurement (the drafting of measurement instruments, calibration, modeling the instrument, etc.) aim to diminish the influence of the human agent on the revealed correlations. Scientific measurement thus is truth-conducive as it ensures that the correlations that we find are as "mind-independent" as we can get.

EE is able to supplement CS for the following reasons: Firstly, EE is a cognitive theory as well, giving an account of how concepts emerge on the basis of correlations between actions and sense input (i.e. correlations revealed through interacting with the world). Secondly, it gives the world an essential place in concept formation processes that goes beyond mere sense inputs: the "real" correlations between actions and sense inputs are learned, and these are not just dependent on our perception, but equally dependent on objective reality. Thirdly, it provides criteria for epistemic objectivity, which is the key element that CS is in need of (see 4.2). It does so by iden-

tifying scientific measurement as one human activity that is designed to identify the most stable and thus most mind-independent correlations that we, as human agents, are able to detect. From a very skeptical point of view, this might still not be "reality as it is by itself", but it will be the most objective truth we can get at. EE thus provides an account that is based on pragmatist intuitions (it is based on our interacting with the world) *and* justifies a certain variant of scientific realism, thereby ensuring epistemic objectivity.

To conclude, we have argued that CS alone is not able to provide a new account of naturalness that would be able to address the traditional philosophical problems that naturalness is taken to solve. However, it provides half of a solution by giving a good account of cognitive sparseness. It can be complemented with EE, providing the missing epistemic objectivity, to jointly offer a truly cognitive account of naturalness that does not have to be agnostic to deep metaphysical questions.

Acknowledgements The research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG) under project number 462340335. We would like to thank our colleagues in Düsseldorf, especially Giacomo Giannini and Markus Schrenk, for most helpful comments and discussions.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Armstrong, D. M. (1978). *Universals and scientific realism* (Vol. 121 & 2). Cambridge University Press.

Bird, A., & Tobin, E. (2024). Natural kinds. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Spring Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2024/entries/natural-kinds/

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New Interdisciplinary Essays* (pp. 141–85). MIT Press.

Callender, C., & Cohen, J. Special sciences, conspiracy and the better best system account of lawhood. *Erkenntnis*, 73(3), 427–447. https://doi.org/10.1007/s10670-010-9241-3

Chang, H. (2022). *Realism for realistic people: A New Pragmatist Philosophy of Science*. Cambridge University Press. https://doi.org/10.1017/9781108635738

Cohnitz, D., & Rossberg, M. (2024). Nelson Goodman. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Spring Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2024/entries/goodman/

Cook, M. (1980). If cat is a rigid designator, what does it designate? *Philosophical studies*, 37(1), 61–64. https://doi.org/10.1007/BF00353502

Crane, J. K. (2021). Two approaches to natural kinds. *Synthese*, 199, 5–6. https://doi.org/10.1007/s11229-021-03328-9

Devitt, M. & Sterelny, K. (1999). *Language and reality: An introduction to the Philosophy of Language*. MIT Press.
Douven, I. (2022). Best explanations, natural concepts, and optimal design. In D. Glass & J. Schupbach (Eds.), *Conjunctive Explanations*. Routledge. https://doi.org/10.4324/9781003184324-12
Douven, I. (2024). Naturalness, scientifc concepts, and the substantivity of social metaphysics. *Philosophia*.
Douven, I., & Gärdenfors, P. (2019). What are natural concepts? A design perspective. *Mind and Language*, (3), 313–334. https://doi.org/10.1111/mila.12240
Gärdenfors, P. (1990). Induction, Conceptual Spaces and AI. *Philosophy of Science*, 57(1), 78–95.
Gärdenfors, P. (1999). Does semantics need reality? In A. Riegler, M. Peschl, & A. von Stein (Eds.), *Understanding Representation in the Cognitive Sciences* (pp. 209–217). Springer. https://doi.org/10.1007/978-0-585-29605-0_23
Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.
Gärdenfors, P. (2003). Inductive reasoning: From carnap to cognitive science. In N. C. Singh (Ed.), *Proceedings of the International Conference on Theoretical Neurobiology* (pp. 93–109). National Brain Research Centre.
Gärdenfors, P. (2007). Representing actions and functional properties in conceptual spaces. In T. Ziemke, & J. Zlatev (Eds.), *Body, Language, and Mind* (pp. 167–195).
Gärdenfors, P. (2008). Reasoning in conceptual spaces. In J. E. Adler & L. J. Rips (Eds.), *Reasoning: Studies of Human Inference and Its Foundations* (pp. 302–20). Cambridge University Press. https://doi.org/10.1017/CBO9780511814273.018
Gärdenfors, P., & Zenker, F. (2013). Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. *Synthese*, 190(6), 1039–58. https://doi.org/10.1007/s11229-011-0060-0
Goodman, N. (1955). *Fact, fiction, and Forecast*. Harvard University Press.
Hempel, C. G. (1945). Studies in the logic of confirmation (I). *Mind*, *54*(213), 1–26.
Jones, J. E. (2023). Locke on Real Essence. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*, Summer Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2023/entries/real-essence/
Khalidi, M. A. (2016). Mind-dependent kinds. *Journal of Social Ontology*, 2(21), 223–46. https://doi.org/10.1515/jso-2015-0045
Kripke, S. A. (1980). *Naming and necessity*. Blackwell.
LaPorte, J. (2003). *Natural kinds and conceptual change*. Cambridge University Press.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343–377. https://doi.org/10.1080/00048408312341131
Lewis, D. (1986). *Philosophical papers, volume II*. Oxford University Press.
Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, *97*(4), 182–197.
Loewer, B. (1996). Humean supervenience. *Philosophical Topics*, *24*(1), 101–127. https://doi.org/10.5840/philtopics199624112
Loewer, B. (2007). Laws and natural properties. *Philosophical Topics*, *35*(1), 313–328. https://doi.org/10.5840/philtopics2007351/214
Massimi, M. (2022). *Perspectival realism. Oxford Studies in Philosophy of Science*. Oxford University Press.
Putnam, H. (1975). The meaning of meaning. *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.
Putnam, H. (1981). *Reason, Truth, and History.* Cambridge University Press.
Putnam, H. (1987). *The many faces of realism. The Paul Carus Lectures 16. LaSalle*. Open Court.
Quine, W. O. (1969). *Natural Kinds. Ontological Relativity and Other Essays*. Columbia University Press. 114–138.
Salmon, N. U. (1981). *Reference and essence*. Princeton University Press.
Sankey, H. (2024). The objectivity of science. *Journal of Philosophical Investigations*, *17*, 45. https://doi.org/10.22034/jpiut.2023.17473
Scholz, S. (2024). Conceptual spaces: A solution to Goodman's new riddle of induction? *Philosophia*.
Schrenk, M. (2016). *Metaphysics of Science: A systematic and historical introduction*. Routledge. https://doi.org/10.4324/9781315639116
Schrenk, M. (2017). The emergence of better best system laws. *Journal for General Philosophy of Science*, 48(3), 469–83. https://doi.org/10.1007/s10838-017-9374-z
Schwartz, S. P. (1978). Putnam on artifacts. *Philosophical Review*, *87*(4), 566–574. https://doi.org/10.2307/2184460
Sider, T. (2011). *Writing the Book of the World*. Oxford University Press.

Slater, M. H., & Borghini, A. (2013). Introduction: Lessons From the Scientific Butchery. In J. K. Campbell, M. O'Rourke, & M. H. Slater (Eds.), *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science*. MIT Press.

Vosgerau, G. (2024). Ecological empiricism. *Philosophia*.

Warglien, M., & Gärdenfors, P. (2013). Semantics, conceptual spaces, and the meeting of minds. *Synthese*, *190*(12), 2165–2193.

Zenker, F., & Gärdenfors, P. (2014). Modeling diachronic changes in structuralism and in conceptual spaces. *Erkenntnis*, 79(81), 1547–61. https://doi.org/10.1007/s10670-013-9582-9

Zenker, F., & Gärdenfors, P. (2016). Continuity of theory structure: A conceptual spaces approach. *International Studies in the Philosophy of Science*, 30(41), 343–60. https://doi.org/10.1080/02698595.2017.1331983

Part II – Publications
Paper C

Conceptual spaces as predictive models

Bibliographic reference. Scholz, S. (2026a). Conceptual spaces as predictive models. Synthese, 207, 237.
DOI. https://doi.org/10.1007/s11229-026-05549-2
Authorship. Sole-authored by Sebastian Scholz.

Synthese (2026) 207:237
https://doi.org/10.1007/s11229-026-05549-2

ORIGINAL RESEARCH

Conceptual spaces as predictive models

Sebastian Scholz[1]

Received: 8 April 2025 / Accepted: 15 March 2026

Abstract
The cognitive science literature suggests that human reasoning is both a similarity machine and a probability engine, but a clear picture of the relationship between these mechanisms is still missing. Recent work combines Conceptual Spaces (CS) – a similarity-based theory about the structure of conceptual content – with probabilistic, or Bayesian theories of reasoning. This synthesis not only reframes the problem but also prompts a broader question – namely, whether the CS account is compatible with Predictive Processing (PP), which is a thriving research paradigm that aims to derive a grand unified theory of cognition from Bayesian principles. If so, then CS can serve as components of predictive models. This paper articulates compatibility constraints, engages critically with ideas from the aforementioned literature, and evaluates their significance for the compatibility question. The literature provides several mechanisms through which similarity and probability may interact in the human mind, including deriving priors from relative region size, distance to prototype, and updating probability distributions over CS using Gaussian mixtures. These mechanisms support what I call partial Marr-horizontal methodological compatibility. This approach is contrasted with the possibility of reducing psychological similarity to probability.

Keywords Conceptual Spaces · Predictive Processing · Bayesian cognition · Psychological similarity · Conceptual representation · Cognitive modeling · Priors · Prototype theory · Probability · Similarity-based reasoning · Bayesian models of cognition · Marr's levels of analysis · Analogical reasoning

✉ Sebastian Scholz
s.scholz@hhu.de

[1] Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Published online: 28 May 2026

1 The puzzle of similarity and probability

This paper investigates whether human reasoning is a similarity machine or a probability engine. The cognitive science literature suggests that it is both, but a clear picture of these mechanisms and their relationship is still missing. Recent work combines Conceptual Spaces (CS) – a similarity-based theory about the structure of conceptual content – with probabilistic, or Bayesian theories of reasoning (Decock et al., 2016; Douven et al., 2025; Poth, 2019; Strößner, 2022; Sznajder, 2017, 2021). This synthesis not only offers new perspectives on the puzzle at hand, but also prompts a broader question. Namely, whether the CS account is compatible with Predictive Processing (PP), which is a thriving research paradigm aimed at a grand unified theory of cognition in which Bayesian principles play a key role. If so, then CS can serve as components of predictive models.

CS addresses conceptual knowledge and reasoning, and thus covers a proper subset of the phenomena targeted by PP (namely, all cognitive capacities). This difference in scope does not count against compatibility; rather, it makes the compatibility question substantive. Either CS fit within the broader PP framework, or at least one account must be revised or abandoned. I argue for the former. Integration is attractive for both sides: CS is primarily an account of the structure of conceptual knowledge at a given time, while PP highlights change over time. And second, some authors worry that PP works well for perception and action, but less so for higher cognitive capacities such as conceptual thinking and reasoning (cf. Friedrich & Fischer, 2024). The present paper, however, does not attempt a full formal integration – that is work for future research. Instead, it articulates compatibility constraints (Sect. 2.2 and 3), engages critically with ideas from the aforementioned literature (Sect. 4), and evaluates their significance for the compatibility question.

The main idea in the literature is to have subjective degree-of-belief functions range over CS. This generates several models of potential mechanisms through which similarity and probability may interact in the human mind.[1] Three of them are discussed below: Deriving priors from relative region size (4.1), deriving priors from distance to prototype (4.2), and iterative updating through Gaussian mixtures (4.3). These models support what I call *partial Marr-horizontal methodological compatibility*. This approach is contrasted with the possibility of reducing psychological similarity to probability (3.3).

The first subsections 1.1 and 1.2, however, aim at understanding the puzzle. An example is developed to get an intuitive grasp of the relevance and explanatory value of similarity and probability. A historical section will give the opportunity to introduce key concepts and situate this paper in a larger context, before Sect. 2 presents the two target theories.

[1] At this point, empirical support for the models is scarce. Accordingly, this paper engages them on a conceptual level. It commits to the psychological reality of the modeled mechanisms insofar as they explain or predict the behavior of real agents.

1.1 A deadly example

Similarity and probability have emerged as cornerstone concepts from the reasoning debate. To get an intuitive sense of their explanatory roles, think about how you recognize something as a death cap mushroom. When processing sensory information from the target object, you retrieve relevant information from memory. This may include past mushroom encounters or semantic knowledge about mushrooms. You then compare the current mushroom to the retrieved representations to make a judgment based on their *similarity*. You may realize, for instance, that this specimen resembles a *typical* death cap (smooth ring at the stem, characteristic bulb at the bottom, etc.), so you conclude it is one of those – and hopefully avoid it. This portrayal matches everyday experience. It is incomplete, however, as such judgments are rarely an all-or-nothing matter. You may grow a mere suspicion when spotting the ring, but upon unearthing the basal bulb you become quite certain. Many researchers would interpret this as a case of probabilistic reasoning, and argue we routinely assign probabilities to our beliefs and update them as new information arrives. Note how the framing of the example suggests that reasoning is both similarity- and probability-based. This is not self-evident, however, but has been debated for decades.

1.2 Historical exposition

The arguably important cognitive task in the example is an instance of categorization – a fundamental capacity that enables humans to group objects, events, and experiences into meaningful categories. Categorization is a paradigmatic form of reasoning, and historical discussions of reasoning are closely tied to the concept of *rationality*. Much could be said about this notion, but the idea is that there are good and bad ways to reason – "reasoning is normatively constrained" (Gärdenfors & Osta-Vélez, forthcoming, p. 3). In the mid-20th century, debates within cognitive psychology and philosophy developed relatively independently of each other. The following briefly introduces those approaches which are directly relevant to our discussion.

1.2.1 Cognitive psychology: heuristics and generalization

Many psychologists of the 20th century were wary of *systematic errors of judgment* which had cast doubt on the reliability of the human mind and its presumed superiority over nonhuman animals. As Schurz and Hertwig (2019, p. 8) point out, such worries rested on a single methodological assumption: "Either individuals behave in accord with the chosen benchmark of rationality, or their cognitive behavior, measured against the benchmark, is irrational [….]." P. C. Wason, for instance, concluded that subjects in his famous card selection task "may have temporarily regressed to more primitive modes of thinking" (Wason, 1969, p. 471) when their behavior violated the rules of classical logic. While deductive logic was long viewed as the paradigm of rational reasoning, other voices began to challenge this view. Most notably, Daniel Kahneman and Amos Tversky shifted the focus to probability and statistics, launching what Schurz and Hertwig (2019, p. 22) describe as "possibly the most influential research program in psychology of the last five decades": *Heuristics and*

Biases. In this context, heuristics are cognitively thrifty shortcuts which are thought to merely approximate probabilistic reasoning, i.e., reasoning that conforms to the axioms of probability theory. The program treated similarity-based reasoning as one such heuristic. In other words, similarity-based judgments were once seen as inferior to probabilistic ones.

One of the most striking examples of how human reasoning departs from probability theory is base-rate neglect (Tversky & Kahneman, 1974). From the axioms of probability theory follows Bayes' Theorem, which provides a principled way to update beliefs in light of new evidence. It states that the posterior probability of a hypothesis given new evidence $P(H|E)$ is proportional to both its prior probability without the evidence $P(H)$ (*base rate*) and the likelihood $P(E|H)$, which is the probability of the evidence given the hypothesis:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

(For our purposes, the denominator can be set aside).[2] It turns out that people systematically underweight prior probabilities in favor of case-specific information. For instance, when told that a randomly selected person is an engineer or a lawyer, but then given an anecdotal personality sketch, participants overwhelmingly base their probability judgments on the perceived similarity between the description and stereotypes of engineers or lawyers – rather than the actual base-rate probabilities. This suggests that similarity may shape reasoning in a way that overrides Bayesian norms, yielding irrational biases.

Contemporary theorists, however, tend to treat both similarity- and probability-based reasoning as potentially rational, or as adaptive depending on the cognitive capacities and reasoning environments at hand. Heuristics are frequently viewed as "fast and frugal" (Gigerenzer & Todd, 1999), and similarity may be a key principle of the "associative" system of reasoning (Sloman, 1996) – an idea that Kahneman himself popularized as part of a dual-process view of cognition (Kahneman, 2011). In a related vein, Tversky (1977) went on to develop an influential set-theoretic account of similarity that highlights how feature overlap and contrast inform judgments and decision-making processes, thus taking similarity more seriously as a principle of reasoning rather than a just source of error. Moreover, similarity has played and continues to play a key role for various theories of concepts such as prototype theory (according to which your DEATH CAP concept is structured around a typical death cap, see Sect. 2.1 and 4.2) – and of course CS.

When it comes to the *combination* of psychological similarity and probability, however, the most relevant cognitive psychologist is Roger Shepard (1962a, b, 1987). After he pioneered geometric representational tools for physical stimuli in the 1960s, which is direct groundwork for CS, he went on to derive the *universal law of generalization* from similarity measures and generalization probabilities: "For a set of

[2] The likelihood of the evidence $P(E)$ is evaluated in consideration of all alternative hypotheses, so this is an abbreviation. However, this term is standardly used for normalization purposes, i.e., to keep the value of the posterior between 0 and 1, so this is not important for our discussion.

stimuli, *[x]* and *[y]*, the empirical probability of an organism to generalise a type of behaviour towards *[y]* upon having observed *[x]* is a monotone and exponentially decreasing function of the distance between *[x]* and *[y]* in a continuous psychological similarity space." (Poth, 2019, p. 9). This pattern is universal in the sense that it holds across diverse stimuli and even species.

Suppose you encounter a mushroom x and learn to avoid it. Should you generalize avoidance to a second mushroom y? The problem is that you cannot know which mushrooms *truly* must be avoided, i.e., which mushrooms belong to the "consequential region" C. Instead, you must "infer the conditional probability that an unknown instance, y, falls under a candidate concept, C, given that a previously observed example, x, falls under C" (Poth, 2023a, b, p. 3). What Shepard found out is that the conditional probability systematically depends on the similarity between x and y. And that similarity is an exponential decay function of distance in a geometric representational model – which is precisely the similarity function used in CS (see Sect. 2.1).

1.2.2 Philosophy: inductive logic

While the CS account is heavily influenced by the psychological debates just outlined, it has roots in philosophy as well. Rudolf Carnap's *Aufbau* (1967) attempts to systematically reconstruct all properties from a basic set of elementary experiences by means of their phenomenal similarity (i.e., the "recollection of similarity" (p. 127) relation Rs).[3] In his later work on *Inductive Logic*, he abandoned the idea of reducing knowledge to a phenomenalistic basis and instead developed formal systems for probabilistic reasoning. A cornerstone feature of these systems are *confirmation functions* which are given by *prediction rules*. They "assign numbers to pairs of propositions: the hypothesis and the evidence; those numbers represent how likely the hypothesis is given the evidence." (Sznajder, 2017, p. 3). Accordingly, the prediction rules have the form of conditional probability statements. In *The Continuum of Inductive Methods* (1952), Carnap introduced a family of prediction rules, known as the λ-rules, which define posterior probabilities based on observed frequencies – modulated by a parameter λ that governs the agent's willingness to learn from new evidence, or 'inductive inertia'.

Carnap's work is relevant for multiple sections of this paper. The concept of inductive inertia reappears in Sect. 4.3. Furthermore, confirmation functions model how scientists *should* reason inductively from past to future observations. That is, in contrast to the psychologists from above, Carnap is *only* interested in normative benchmarks of reasoning. By contrast, the current paper is empirically oriented, i.e., interested in how real cognitive agents think and reason – as is the CS account. This does not mean that normative considerations are irrelevant – see Sect. 3.1 for a discussion. Most importantly, in his latest and least received writings on inductive logic, Carnap introduced the notion of *attribute spaces*, which is a predecessor of

[3] Carnap's treatment of similarity was later criticized by Goodman (1972), however, I presume here that Goodman's criticism does not affect both Tversky's set theoretic and Gärdenfors' geometric account of similarity – see Decock and Douven (2011) for an argument to this conclusion.

CS.[4] Roughly speaking, attribute spaces provide the semantics for the symbols and propositions that confirmation functions operate on within *applied* – as opposed to *pure* – inductive logic (see Gärdenfors, 2011, Sect. 7 for reference). They "lay down requirements that primitive attributes and relations must fulfill to be admissible as primitive concepts in an object language […]" (Carnap, 1971, p. 70). Crucially, Carnap defined probability distributions on top of attribute spaces, which the literature to be discussed in Sect. 4 takes inspiration from.

To close the historical section, note that, while not relying on Bayes' rule, "it is no longer unusual to consider inductive logic as a branch of Bayesian epistemology, concerned with formal modeling of inductive reasoning as going from a set of assumptions — statistical assumptions, the choice of a conceptual framework, etc. — to conclusions that follow from them by virtue of the formal system." (Sznajder, 2017, p. 14). And while Tversky & Kahneman do not reference Carnap directly, his work clearly contributed to the conceptual groundwork of the heuristics and biases debate.

2 Two theories

We can now introduce the two theories whose relationship will occupy the remainder of the paper. I begin with CS, turn to PP, and then reflect on their compatibility. The aim is to clarify what it would mean to treat CS as predictive models, before discussing in detail how similarity and probability might interact in the human mind.

2.1 Conceptual spaces

Concepts are traditionally conceived as the building blocks of thought (Margolis & Laurence, 2021). Within cognitive psychology, they are postulated to explain higher cognitive capacities such as categorization and reasoning. The key claim of the CS account is that concepts, or more precisely *conceptual contents*, have a *geometric structure* (Gärdenfors, 1990, 2000).[5] To make this vivid, think of semantic knowledge as organized into internal "landscapes". From behavioral data such as similarity judgments, we can calculate geometric models, or "maps" that capture the topology of these landscapes (e.g., by *multidimensional scaling*, cf. Shepard, 1962a, b and Gärdenfors, 2000, Sect. 1.7). Such models are thought to explain and predict the observed behavior. They are called "similarity spaces", as the distance between points represents the (dis-)similarity of objects. A *conceptual* space is what you get when you partition a similarity space into regions which are interpreted as concepts. A standard example is the property of *being green*, which is identified with a region

[4] Despite the similarities between CS and attribute spaces, Gärdenfors contrasts his own take on reasoning with the logicism characteristic especially of Carnap's early philosophy. His point is that "we have to go below language" (Gärdenfors, 2011, p. 3) to understand reasoning, while Carnap and the Vienna Circle operated on a purely symbolic, or propositional level.

[5] Several versions of CS can be found in the literature. I focus on Gärdenfors' account here because it is well developed and engages in philosophical discourse.

in a three-dimensional "visual domain" ("domain" meaning the dimensions come as a package).

This approach has several advantages. Its formalism is well-developed, and Gärdenfors' books and subsequent essays explicitly connect the framework to empirical literatures and outline how to test his theory against data. A fair amount has happened in this regard, although I cannot review the empirical literature here.[6] A further advantage is that it covers prototypes – the best representatives, or most typical instances of concepts. Prototypes are well known to affect cognitive processing, and accounting for such prototype effects is desirable for a theory of concepts (Hampton, 2006; Rosch, 1973). Gärdenfors (2000) views prototypes as the centers of gravity of *convex* conceptual regions (Sect. 3.8). "Convexity" means that if two points lie in a given region, all points between them lie in the region as well. Given a set of prototypical points, a partitioning of the space into convex regions can be generated by *Voronoi tessellation*, where every point is allocated to the closest prototype. Convexity is an expression of a high degree of topological cohesion. Accordingly, the points within a convex region exhibit a high degree of mutual similarity. Following a line of thought by Quine (1969), and considering principles of cognitive economy, this is taken to indicate the *naturalness* of a concept – although "cognitive sparseness" might be a better label, since CS targets the joints of thought rather than the objective joints of nature (see Scholz & Vosgerau, 2025).

The philosophical merit is to get rid of gerrymandered concepts such as Goodman's (1955) property of *being grue*. In that sense, CS aligns with Carnap's attribute spaces, which also determine which concepts are permissible for reasoning. The convexity criterion, however, is a unique feature, and as Gärdenfors points out, Carnap "does not take […] metric as a primitive notion" and "never leaves the ideal that all knowledge should be expressed symbolically […]" (2011, p. 10). Another interesting point of comparison is semantic internalism, which is the idea that the meanings of words (or mental states) are a matter of internal psychological states, rather than external worldly affairs (see Putnam, 1975). Both Carnap and Gärdenfors are internalists, but the latter's position is nuanced. Gärdenfors advocates for cognitive semantics, meaning he takes language to refer to internal conceptual structures. However, these structures are thought to develop in *interaction* with the world:

> Via successful and less successful interactions with the world, the conceptual structure of an individual will adapt to the structure of reality. It must be emphasized, however, that this does not entail that the conceptual structure represents the world. Gärdenfors 2000, p. 156.

This passage expresses *pragmatism* with evolutionary overtones – a Neo-Kantian view, as cognitive systems lack access to the distal world in itself and engage instead

[6] For a selection, I refer the interested reader to work on Shepard's law (Shepard, 1987), RSA (Kriegskorte & Kievit 2013), optimal partitions of color space (Douven & Gärdenfors 2019, Regier et al., 2007), spatial codes (Bellmund et al., 2018; Constantinescu et al., 2016), graded membership (Douven, 2016), category-based induction (Osta-Vélez & Gärdenfors 2020), robotics (Cubek et al., 2015), and convexity in deep networks (Tětková et al., 2025).

with what is proximally relevant for a survival. Gärdenfors' position is best characterized as weak realism (see Scholz, 2024 for a discussion).

We will return to this point later, but the key takeaway of this section is that CS explains cognitive capacities that are associated with conceptual knowledge by appealing to psychological similarity – modeled as spatial distance.

2.2 Predictive processing

By contrast, PP has broader ambitions. While much work applies the paradigm to specific cognitive capacities, leading figures present it as a "grand unified theory" of the brain and cognition (Clark, 2013, 2015; Friston, 2002, 2005, 2008, 2010, 2012; Howhy, 2013, 2016). Accounts under the PP umbrella vary considerably. As Sprevak and Smith (2023, p. 2) caution:

> PP should be understood, not as a fully articulated model of cognition, but as a broad research program in computational neuroscience with a set of evolving commitments that are only gradually being uncovered and agreed upon. It would be a mistake to try to state now what PP, as an umbrella framework under which many distinct models could fall, is and is not committed to.

Even so, several interconnected core tenets recur in the philosophical discussion of PP.

(T1) *Cognition is driven by prediction error minimization (PEM).* To appreciate this insight, imagine you are a programmer who wants to compress an image. Obviously, you do not want to store every single pixel – that would be no compression at all. Luckily, you don't need to, since the value of each pixel predicts the value of its neighbors. Thus, you only need to account for the pixels where something *unexpected* happens, which yields an efficient compression strategy. This strategy, which also underlies JPEG-compression, is an instance of *predictive coding*.[7] And the idea behind T1 is that the brain is a predictive coder, too. It constantly generates top-down expectations, and processes bottom-up information only when prediction error occurs, that is, when incoming information doesn't match the expectations: "Transposed to the neural domain, this makes prediction error into a kind of proxy for sensory information itself." (Clark, 2013, p. 183). Minimizing the error comes down to navigating the world as efficiently and successfully as possible. This view changes our understanding even of fundamental cognitive processes such as perception, which, if correct, is more of a "controlled hallucination" (Clark, 2015, p. 14) than everyday experience would suggest.

(T2) *The brain achieves PEM by utilizing hierarchical generative models*. Of course, the nervous system is not monolithic but organized into nested, hierarchically related sub-systems. Early visual processes (e.g., retinal and V1 circuitry) han-

[7] Note that this term is sometimes used synonymously with PP; at other times, it takes on more specific meanings (e.g., Sprevak & Smith, 2023 introduce it as a "prominent PP model of perception"). Here, it refers to methods of data compression and transmission that focus on encoding only the differences between expected and actual values.

dle relatively simple features such as edges, while later stages encode increasingly abstract, multimodal structure – up to "a wealth of multimodal associations and […] a complex mix of motoric and affective predictions" (Clark, 2015, p. 14). In PP, subsystems actively predict the inputs from the environment or downstream systems by modeling their sources.[8] On the agent-level, this amounts to generating the worldly causes of the sensory input.

A central idea is that agents can either adjust their models or *act* to make the world match their predictions. Say, you are searching for mushrooms by deploying nested top-down models of what forest floors, typical death caps, etc. should look like. Upon spotting something light green, you don't just stand there and compute, but *move* towards the object, dig out the basal bulb to make sure it's a death cap, and so on. Crucially, this whole scenario is about generating and testing hypotheses. It can thus be characterized as an inferential process that blurs the very distinction between perception and action:

> […] it now seems likely that perception and action are in some deep sense computational siblings […]. This deep unity […] emerges most clearly in the context of active inference, where the agent moves its sensors in ways that amount to actively seeking or generating the sensory consequences that they (or rather, their brains) expect. (Clark, 2013, p. 186).

Accordingly, (T3) *agents engage in active inference*. While the tight coupling between perception and action is standard in PP, views differ on how tightly the mind is coupled with the body and the world. Jacob Howhy, for instance, emphasizes internalism and methodological skepticism: "PEM reveals the mind to be inferentially secluded from the world, it seems to be more neurocentrically skull-bound than embodied or extended, and action itself is more an inferential process on sensory input than an enactive coupling with the body and environment." (2016, p. 259). Andy Clark, by contrast, embraces externalism:

> […] 'inference', as it functions in the PP story, is not compelled to deliver internal states that bear richly reconstructive contents. It is not there to construct an inner realm able to to [sic!] stand in for the full richness of the external world. Instead, it may deliver efficient, low-cost strategies whose unfolding and success depend delicately and continuously upon the structure and ongoing contributions of the external realm itself […]. (2015, p. 191).

He combines this view with *embodiment*, treating the body as a constitutive part of cognition. Gallagher and Allen (2018) push further toward enactivism, some versions of which aim "to replace all representational explanations of cognition with embodied, interactive explanations" (Shapiro & Spaulding, 2024, Sect. 2.4). Thus,

[8] To the uninitiated, it may sound suspicious to speak of neural sub-systems as "modeling" or "actively predicting". However, this way of speaking serves as a shorthand for sub-personal processes such as "automatically deployed, deeply probabilistic, non-conscious guessing that occurs as part of […] complex neural processing routines […]" (Clark, 2015, p. 2).

while (T3) implies that the relationship between system and world is active, it does not imply that it is direct in the sense of externalism. Moreover, note that the notion of active inference can take a number of different meanings in the literature (e.g., as a specific decision-making algorithm in Sprevak & Smith, 2023).

(T4) *PEM approximates Bayesian inference under uncertainty.* In traditional Bayesian reasoning, a system "[…] tests a prior belief or hypothesis (the "prior" expressed as a probability distribution of some event or parameter) against some incoming data in order to affirm the prior or update it according to Bayes rule." (Gallagher & Allen, 2018, p. 2629). In PP, "[m]ismatches (i.e., prediction errors) between the prediction [prior] and the data [evidence] are then propagated 'forward' in the system where they serve to further refine the original hypothesis. This process puts into effect a continual scheme of updating 'empirical priors' with sensory evidence" (ibid.). In this minimal sense, PP is Bayesian: It characterizes perception and learning as the ongoing adjustment of probabilistic expectations in the light of sensory evidence.

However, PP should not be identified with Bayesian reasoning. A characteristic feature of PP is the hierarchical organization of expectations: what functions as a prior at one level is itself shaped by learning at higher levels (this is formalized by the "empirical Bayes" account from statistics, cf. Robbins, 1992). Moreover, PP explicitly acknowledges that biological systems are not ideal reasoners: Sensory evidence varies in reliability, and posterior computation is generally intractable for biologically realistic models. Accordingly, PP theorists agree that inference must be sensitive to input uncertainty or noise (cf. Knill & Pouget, 2004) and must rely on approximation methods. In this sense, PEM is better understood as a family of biologically plausible procedures that approximate Bayesian inference, rather than an instance of Bayesian inference itself.

Accounts diverge, however, in how far they go in specifying the mathematics and implementation of such approximation. Clark accepts T4 but does not commit to any particular algorithm for computing posteriors. Indeed, he "leaves plenty of space for other ploys and strategies to coexist with the core PP mechanism" (2013, p. 243). Friston and Howhy, by contrast, embrace the *free energy formulation* of PP, which "represents a kind of 'maximal version' of the claims concerning the computational intimacy of perception and action" (Clark, 2015, p. 306). Free energy is "an information-theoretic isomorph of thermodynamic free energy in a system's exchanges with the environment" (ibid., p. 305), and minimizing it is taken to be a general requirement on biological systems.

Formally, this proposal connects PP to *variational Bayes*. This is a family of techniques for approximate Bayesian inference in which an intractable posterior is replaced by a tractable approximation, and the quality of this approximation is evaluated via Kullback-Leibler (KL) divergence between the two distributions (cf. Blei et al., 2017). Free energy enters the scene because "an agent that aims to approximate Bayesian inference […] can always be described as minimizing variational free energy." (Sprevak & Smith, 2023, p. 6). The idea is that biological systems must keep surprisal low (where surprisal is the improbability of sensory input under the system's model) in order to maintain themselves. However, surprisal cannot be computed directly because this would require marginalizing over hidden causes. Variational

free energy can be written as surprisal plus KL divergence between the approximate posterior and the true posterior. Minimizing it therefore indirectly minimizes surprisal (cf. Friston, 2005, p. 821). In statistics and machine learning, this provides a tractable objective for fitting the approximation, despite computational intractability of the target posterior (cf. Yellapragada & Prakash, 2019). The free energy formulation of PP then presents this strategy as a general principle of biological self-organization.

It is not the aim here to adjudicate between these positions or to survey the empirical literature – though PP is increasingly prominent in psychology and supported by a growing body of neuroscientific evidence.[9] The takeaway is that *CS qualify as predictive models if they are used to generate predictions as specified by tenets T1–T4.* We can treat these tenets as constraints that CS must meet to be compatible with PP. The next section approaches the issue from the opposite direction by specifying constraints that PP accounts must satisfy to be compatible with CS.

3 Compatibility

At first glance, the two accounts differ markedly: CS concerns conceptual knowledge, whereas PP targets cognition at large. CS is similarity-based; PP is probability-based. Regarding reasoning, CS emphasizes criteria of cognitive sparseness, while PP highlights Bayesianism and the coupling of perception and action. One might even think that CS offers a relatively static "sandwich" view of cognition – "where cognition forms the content, perception the input and action the output of the system" Poth, 2025, Sec. 3.2), since it models semantic knowledge at a time rather than detailing how such knowledge is learned and updated. Even so, there are good reasons to expect a degree of compatibility. While Gärdenfors does not detail how CS are updated, we have seen that *interaction* lies at the heart of his philosophy of science, metaphysics, and epistemology, which sits well with active inference. Moreover, the literature discussed in Sect. 4 integrates CS with Bayesian reasoning, which strongly supports compatibility.

"Compatibility", however, is multivalent. Without surveying the extensive literature on relations between theories, I adopt some basic, pre-theoretical distinctions. Broadly, theories may align – or fail to align – in terms of their *goals*, *ontology*, and *methodology.* This list is non-exhaustive and intended to structure discussion rather than settle the matter. In what follows, I will derive compatibility constraints from these dimensions. The section on methodology will enable us to see that the current literature demonstrates a specific kind of methodological compatibility.

3.1 Goals

Differences in scope (conceptual knowledge vs. cognition broadly) may of course count as a misalignment of goals, but, as argued in the introduction, this does not undermine compatibility. More pressing and more complex is an issue that has been labeled "normative-descriptive interplay" (Strößner, 2023, forthcoming).

[9] See Hodson et al. (2024) for a recent review.

Gärdenfors (2000) highlights that his aims are descriptive: to explain and predict the behavior of cognitive agents and to guide the construction of artificial agents. PP, by contrast "is based in probability theory – Bayesian epistemology – which is normative because it tells us something about what we should infer, given our evidence" (Howhy, 2013, p. 15). This apparent conflict, however, does not threaten compatibility.

The first thing to point out is that PP is not identical to Bayesian reasoning: Clark does not even mention normativity in his 2015 opus magnum, and PP agents are not ideal reasoners, but merely approximate ideal reasoning by minimizing prediction error. There is a second point, however, which runs deeper. Bayesians think that ideal agents should reason in accordance with probability theory. If this is conceived as *instrumental rationality*, which involves conditional normative statements – what is good or rational, *given* some benchmark – then:

> […] normative epistemology is a branch of engineering. It is the technology of truth-seeking […] There is no question here of ultimate value […]; it is a matter of efficacy for an ulterior end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed. (Quine, 1986, pp. 664–665).

Statements about belief updates cohering with axioms may not be empirical, but are still descriptive. The main alternative is for Bayesians to pursue *normative rationality*, which involves categorical normative statements. Declaring Bayes-optimal reasoning as an absolute, goal-independent norm, however, sneaks in a value judgment about what the correct benchmark is – an is-ought fallacy (Elquayam & Evans, 2011).[10] If this is correct, then Bayesianism either turns out more descriptive than is commonly appreciated, or turns out fallacious.

Furthermore, one might argue that the CS account is normative after all, as Douven & Gärdenfors's (2019) elaboration appeals to *optimal* design. However, the authors explicitly frame this as an "engineering task […] where it is given that the system must succeed in a competition for scarce commodities" (p. 6), which places them squarely in the instrumental rationality camp. Thus, the CS account should be understood as a descriptive account rather than a normative or even "hybrid" (cf. Strößner, 2023) one.

Yet, it would be hasty to demand of PP accounts to be descriptive in order to be compatible with CS because there are important differences within the instrumental rationality camp. The crucial point is that, while some Bayesians are only interested in ideal reasoning, the CS account retains a strong empirical orientation. Optimality criteria and constraints are only proposed insofar as they capture psychological and simulation studies. Accordingly, the compatibility constraint to be derived here is (C1) *basic empirical orientation*. In contrast with pure Bayesianism, PP accounts can generally be assumed to fulfill C1 because they map how *real* agents approximate ideal ways of reasoning.

[10] By contrast, Schurz and Hertwig's (2019) account is normative in the sense of justifying reasoning benchmarks by grounding them in cognitive success.

3.2 Ontology

Both theories posit cognitive agents that reason about worldly affairs. CS, however, posits conceptual representations with certain topological features that are foreign to PP. This may conflict with radically enactivist accounts due to their anti-representationalism. But positions in the "representation wars" (Gallagher et al., 2022, p. 1010) are nuanced and rarely radical, so this must be evaluated on a case-by-case basis. As a second compatibility constraint, consider: (C2) *openness to minimal (geometric) representationalism in conceptual processing*. Some cognitive capacities that are associated with conceptual knowledge must be at least partly explained by CS representations. *Prima facie*, the accounts presented here can be assumed to fulfill (C2) because they are not radically enactivist.[11] Nevertheless, (C2) is important because some lesser-known PP accounts are explicitly anti-representational (e.g., Bruineberg & Rietveld, 2014; Facchin, 2021; Kirchhoff & Robertson, 2018).

With respect to fundamental ontology and the mind-world relation, Gärdenfors' stand in the realism debate aligns differently with different PP accounts. Cognitive semantics is arguably a form of internalism, which first suggests a high degree of compatibility especially with Howhy's account – recall the concept of inferential seclusion. He would certainly agree with Gärdenfors that representations do not refer directly to external entities, i.e., that semantic externalism is false. On the other hand, the evolutionary pragmatism of CS does underline interactive engagement with the world and can hardly be said to view the mind as secluded. However, since Gärdenfors introduces this idea mainly to avoid relativism (see 1990) and never works out its details, it is unclear whether a deep conflict arises. Complicating matters, it can be quite difficult to pin down the ontological commitments of PP accounts. But "[i]f Bayesian models are not to be understood realistically, philosophical controversies about the status of representations in Bayesian models, or their bearing on externalist (or internalist) arguments about mental states, are pointless." (Colombo et al., 2021, p. 188). Because of the opaqueness surrounding these issues, there is no clear compatibility constraint to be derived from the discussion of fundamental ontology. Moreover, (C2) is formulated as a mild constraint that is neutral on the ontology of concepts: Explanations must involve geometrically structured conceptual representations, whether they are mental objects, abilities, or something else.

3.3 Methodology

Regarding methodological compatibility, there appears to be a clear divide at play: CS works with similarity-based models, whereas PP relies on probabilistic models. One problem is that there is a great variety of such models, developed for a great variety of cognitive phenomena. It is thus hard to discuss methodological compatibility with any generality. Insight can be derived from looking into a specific account,

[11] This is true even for Gallagher et al. 2022 who characterize their 'representation-ontological' position in the cited footnote as follows: "hands, bodily movement, goggles (or artifacts and tools more generally), various ecological factors, material affordances, agentive history, extra-neural factors, etc., can[not] simply be reduced to stored representations." This is not necessarily in conflict with CS because it does not rule out that conceptual representations have explanatory value.

however: Tenenbaum and Griffiths's (2001, T&G) Bayesian model of perceptual generalization. Recall that Shepard captures generalization through a universal law based on probability over geometric similarity. T&G, by contrast, describe the same behavior with a purely Bayesian model. More specifically, the authors attempt to *subsume* the *algorithmically* specified similarity models by Shepard and Tversky on a *computational* level. In Marr's (1982) scheme, the computational level specifies *what* is being computed and *why*. The algorithmic level specifies *how* the problem is solved – the representations used and the processes that transform them. In other words, by shifting to a different level of explanatory analysis, T&G aim to *reduce* similarity to probability in a mathematical modelling sense.[12] Successful reduction would arguably speak for the compatibility of CS and PP, even if the latter is not identical to Bayesian reasoning.

In Shepard's model, reasoners infer the conditional probability that a single stimulus belongs to a consequential region within a geometric similarity space. In T&G's model, reasoners maintain a probability distribution over a hypothesis space H of hypotheses h specifying which consequential region is responsible for one or multiple stimuli. The distribution is updated via a modified Bayes-rule which replaces the likelihood by the *size principle*. T&G compute the generalization function:

$$P(y \in C|x) = \sum_{h:y\in h} P(h|x)$$

which yields the exponentially decreasing curve known from Shepard. The summands are "the probabilities $p(h|x)$ of all hypothesized consequential regions that contain y" (T&G, p. 631). They are posteriors which are calculated with the modified Bayes-rule: For n observations $X = \{x_1, \ldots, x_n\}$, T&G set the likelihood

$$P(X|h) = \frac{1}{|h|^n}$$

The idea behind this formula is that "smaller hypotheses receive higher likelihoods than larger hypotheses, by a factor that increases exponentially with the number of examples observed" (T&G, p. 634). That the number of examples matters should be intuitive. The insight of the size principle is that a small region that contains some set of observations is more likely to produce that set under random sampling than a large region that contains the same set. In other words, the size principle concerns the information value of the hypotheses, and tells us why it is rational for learners to prefer smaller hypotheses.

Poth (2019) critiques T&G's account for its *semantic opaqueness*:

> [It] leaves open how the size of a candidate category should be measured, and its semantic content interpreted. One reason for its semantic opaqueness is that [it] loses the explicit mentioning of the evidence […]. This also makes it difficult to identify the degree of confirmation of a belief by an example because it

[12] For an account that takes similarity to be fundamental, see Strößner (2024, conference talk).

> cannot be compared how far and by which measure the contents of terms [X] and [h] align. More conceptually, it is not clear what the belief in the Bayesian model that is assigned a probability value is actually about. (Poth, 2019, p. 16).

This notion is distinct from – but related to – the hotly debated topic of *AI opacity*, where the central issue is that AI systems are black boxes. To my knowledge, Facchini and Termine (2022) are the only authors from this debate who offer a fitting characterization: "[…] the format used to store and manipulate information prevents users from giving it a meaningful interpretation […]" (p. 12) – either because there is no well-defined semantics, or the users lack the skills to understand it, or it is inadequate for the context. Usually, opacity concerns the user's *access* (e.g., to training data) or the *link* between model and target system.

T&G do not explicitly define a semantics for their probabilistic model. Consequently, *rather than demonstrating that probability is more fundamental than similarity*, *they end up presupposing similarity while obscuring its role*: Hypothesis size depends on a structured space of possibilities H where some hypotheses h subsume others. That structured space functions like a similarity space, even if the model does not explicitly define one – or at least this is Poth's main point.[13]

The issue is not confined to T&G, but an instance of a common thread: "Bayesian inference itself is conceptually trivial, but strong assumptions are often embedded in the hypothesis sets and the approximation algorithms used to derive model predictions, without a clear delineation between psychological commitments and implementational details" (Jones & Love, 2011, p. 169). Opaque semantic assumptions are a direct result of the computational focus of Bayesian models which leave representational format and psychological content unspecified.

While this result speaks against a specific form of reduction, it does not imply that CS and PP are incompatible. First, note that CS is best understood at the algorithmic level: it specifies semantically interpretable representational primitives (quality dimensions, metrics, etc.) and admissible procedures over them by which tasks like categorization or induction are carried out. CS thereby tells us how information is represented and transformed while remaining agnostic about the physical substrate (hence not implementational)[14] and not, by itself, fixing the rational objective (hence not purely computational). In other words, compatibility might arise from the accounts operating at *different* levels of explanatory analysis. Second, PP is not identical to Bayesian inference, and PP models exist at all three levels (see Sprevak & Smith, 2023). Given the previous discussion, the most straightforward variant would be for CS to realize the computational goals of PP algorithmically. In its current form, however – despite highlighting the philosophical significance of interaction – the CS-account is static in the sense that we know little about how the structure and

[13] Poth's critique is based on the requirement that explanatory models in cognitive science should tell scientists how cognitive content produces behavior. T&G fail to do so because (i) they opaquely presuppose similarity, (ii) they run into the symbol grounding problem, and (iii) they fail to rule out gerrymandered categories. Poth uses the CS account as a remedy. See Scholz (n.d.) for a detailed analysis of her critique and solution.

[14] The implementation level is the third, lowest level and concerns the physical realization of the algorithm (e.g., by neurons). Higher levels are thought to be multiply realizable – but constrained – by lower ones.

conceptual partitioning of the spaces might change over time.[15] Since PP centers on how world models should be updated, CS cannot yet realize PP's computational goals algorithmically. A natural direction for future work is to study how people's spaces update with new information and whether such updates approximate optimal solutions in Bayesian networks. Other level-combinations may also be fruitful. The present paper, however, surveys existing proposals and evaluates their relevance to compatibility; that literature is best understood as combining elements of both frameworks within a single level, whether computational or algorithmic.[16]

4 Horizontal integration

To recap, CS accounts are compatible with PP if they accord with tenets T1 – T4. PP accounts are compatible with CS, if they satisfy C1 and C2, and if they *accommodate the method of using geometric-similarity-based models* (call this C3). Regarding C3, the issue of semantic opaqueness speaks against T&G's reductionism, but different combinations of Marr-levels are viable. What follows discusses mechanisms that integrate similarity and probability horizontally at the same Marr-level.

Note that the focus is on how similarity and probability might interact in real human minds (C1, empirical orientation). As a contrast, consider Brössel (2017) who presents an epistemically oriented model of categorization. He discusses the long-standing philosophical debate on whether the transition from (non-conceptual) perceptual experience (PE) to perceptual beliefs (PB) is rationally justifiable. He formally demonstrates that Bayesian inference from PE-contents modeled as point observations in similarity spaces to PB conceived as categorizations on conceptual regions is feasible. In other words: the combination of CS and Bayes provides the desired normative reconstruction.[17] By demonstrating that such combination is philosophically fruitful, he provides direct support for the thesis of non-reducibility – because probabilistic reasoning is indispensable for the justification of the transition between similarity-based contents.[18] For a discussion of the rational justification of PE and PB in PP (without similarity, however), see Gładziejewski (2021).

4.1 Priors by relative size (PRS)

Decock et al. (2016) propose deriving priors from the relative sizes of conceptual regions – a move directly inspired by Rudolf Carnap. In Carnap's writings, priors serve as inputs for confirmation functions, which tell us what hypotheses one should

[15] This is not to say Gärdenfors doesn't discuss issues like concept learning or scientific theory change, but it is fair to say that these aspects of his theory are algorithmically underspecified.

[16] A remark by Douven et al., 2025 (see below) indicates that they take their model to operate on the computational level. However, they settle on mathematical procedures and representational formats, so this might be contested. It is also possible that the distinction between computation and algorithm isn't always clear, but this discussion goes beyond the present paper.

[17] Brössel (2017, p. 733) mentions "instrumentalist, a priori arguments for Bayesianism", so his reconstruction presupposes certain normative benchmarks such as coherence.

[18] I am grateful to an anonymous reviewer for highlighting this point.

believe, given a conceptual framework and a single observation. In PP, such priors would serve as inputs for iterative Bayesian updating with multiple observations. A CS account of the latter will be presented in Sect. 4.3.

Say, you draw a colored ball from an urn. How likely are you to draw a certain color, e.g., green? The *principle of indifference* (PoI) states that, absent further information, you should treat all options as equally likely. The problem is how to carve up the option space: green vs. non-green (prior = 0.5), green vs. red, blue, and pink (prior = 0.25), and so on. This is where Carnap makes use of his attribute spaces:

> Once we assume that the first observed object can have any of the qualities represented in the attribute space, and there is no reason to favour any of them, the prior probability can be thought of as being uniformly distributed over the set of points of the space. From that it follows that the prior probabilities for predicates will be equal to the sizes of the regions that represent them. (Sznajder, 2021, p. 469).

The idea works because a uniform density over points is invariant under how the space is partitioned. Decock et al. (2016) render it mathematically precise in CS (Carnap's treatment is rather programmatic) and develop an account of static rationality, i.e., of what is rational to believe at a specific time. Whereas Carnap's approach allows a free choice of attribute spaces relative to the scientific purpose at hand, their geometric PoI operates on CS derived from similarity and typicality judgments of real agents. These are said to be "objective" and therefore suitable for modeling static rationality, which may be criticized by pointing out that Gärdenfors' evolutionary pragmatism is not sufficient to establish a stable connection to the world that would ground robust ontological or epistemic objectivity (cf. Scholz & Vosgerau, 2025).

Formally, to calculate the size (measure) of a region $T \subseteq S$ in a continuous n-dimensional space S, define the indicator function:

$$\mathscr{I}_T(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } (x_1, \ldots, x_n) \in T \\ 0, & \text{otherwise} \end{cases}$$

The Lebesgue-measure $\mu(T)$ of T is then:

$$\mu(T) = \int_S \mathscr{I}_T(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

In other words, the probability of a region T is the integral of the pointwise density over T. To obtain a probability rather than a raw measure, normalize by dividing by $\mu(S)$. Decock et al. also refine the measure to accommodate prototypes, graded membership, and fuzzy boundaries; those details are not essential here (see also Douven et al., 2013).

One may wonder whether the idea behind PRS is sound. Consider: if the size of the *green* region in my visual domain carried information about the probability of encountering green objects in the future, then that region should change size as one observes many green objects. Prima facie, this is implausible, since region size

and partitioning depend on considerations other than frequency – e.g., cognitive efficiency, informativeness, and parsimony (Douven and Gärdenfors, 2019). Moreover, the points in CS are not real observed objects, but possible objects! They are individuated by their position along the dimensions, which in turn represent the various qualities that objects can have. Thus, in a certain sense, it is not possible in CS to observe two different objects with the same exact qualities. The point is nicely illustrated by Gauker (2007, p. 333) in the context of persistence:

> Suppose I judge that John is a college student. […] According to the similarity space theory of concepts, my resulting judgment, or belief, that John is a college student may be represented as consisting of my having a point representing John in the college student region of my similarity space. But suppose that subsequently I learn that I was mistaken; John is not a student after all. […] Whereas earlier I had a point representing John in the college student region of my similarity space, I will now have a point representing John outside of the college student region of my similarity space. Here is my question: In what sense can the earlier point in my college student region represent the same thing as the later point outside of my college student region? There is no question of these two points being the same point; two points in different regions cannot be the same point. At best we can say that these two points represent the same object, namely, John. How so?

This might suggest that information about the individuation and probability of occurrence of real objects is stored completely separately from CS. But that conclusion may be too quick. Regarding persistence, an available reply is that represented individuals are only mapped on CS, and the mapping can change. And even conceding that regular points do not carry probability information, CS come with a rich structure that may be utilized – Sect. 3.2. considers points with a special status, namely prototypes.[19]

To be clear, none of this implies that the geometric PoI is unjustified. Carnap's theoretical motivation stands. However, PRS seems apt only under the idealized assumption that no other information is available. This may be reasonable for normative models of reasoning, but is less suitable for describing real cognitive agents. Expectations and previous knowledge about frequencies enter even tasks such as sampling colored balls (e.g., green balls will be expected more than glaucous ones). Or as Sznajder (2017) puts it:

> Only in the idealized case can there be an agent whose conceptual framework is given prior to possessing any other knowledge or presuppositions about the features of the world. In real life rather we encounter an ever-descending chain of conceptual dependencies, with one investigation after another shaping our

[19] By refining their account to cover prototypes and degrees of membership in regions and with fuzzy boundaries, Decock et al. (2016) already make use of that structure. However, their overall approach for deriving priors is based on relative region size, which is different from what is discussed in the next section.

> conceptual frameworks, which in turn are used again and again to gather more knowledge. (47–48).

Thus, PRS is not a promising account of how similarity and probability interact in real-life human reasoning.

4.2 Priors by distance to prototype (PDP)

And yet, intuitively it makes sense to think that our conceptual repertoire shapes our expectations and predictions about the world. I would be surprised to find a purple death cap because my conceptual knowledge tells me that death caps are typically light green. This line of thought suggests a connection between probability and prototypicality. Considering that features of biological systems are normally distributed in populations, which prototype concepts might have evolved to mirror, the intuition that what's typical is probable gets even stronger (cf. Schurz, 2012).

Thus, one may hypothesize that cognitive agents derive priors for Bayesian reasoning from geometric information on prototypes (engage in PDP). But here we must proceed carefully and be precise about what typicality information is correlated with what probability information.

Here's something that doesn't work: The typicality of some member M of category C does not license conclusions about the probability that $M \in C$. Highly untypical members can still belong to the category with very high certainty – in short: untypical death caps are just as deadly. Likewise, inferring real-world frequency from subjective typicality is ill-advised. One might think that sharks typically kill people, although fatal shark attacks are exceedingly rare (most species are small and feed on mollusks or crabs).[20] Such misalignments are unsurprising in light of the heuristics-and-biases literature.

Now for something more promising. Prototypicality clearly carries information about what properties to expect from members, as members *inherit* them from their prototype:

> […] our expectations about the world mirror aspects of the organization of our background knowledge. For instance, if we know that Maria bought a new pet, we would expect it to be a typical one like a dog or a cat, and, to a lesser extent, a bird or a fish. We would be surprised if we then learn that Maria's new pet is a lizard since it is rather rare to think of these animals within the scope of domestic pets. (Osta-Vélez & Gärdenfors, 2022, p. 82).

For empirical and philosophical work on inheritance in a frame-theoretic model of conceptual knowledge, see Strößner et al. (2021) and Schuster (unpublished). In the cited paper, Osta-Vélez & Gärdenfors develop a CS account of *expectation orderings* in non-monotonic reasoning that is based on distance to prototype. In non-monotonic logic, premises are considered differently defeasible, which is modeled here by a

[20] I am grateful to Igor Douven, Gerhard Schurz, and Matías Osta-Vélez for their examples and insights. Without them, this section would be in worse shape.

"classical consequence relation plus an expectation-based ordering of formulas" (p. 77). However, the model devised by the authors does not take number of occurrences into consideration, as "[p]robabilistic models will not give the right results for expectation orderings since some property that is probable may be very atypical." (p. 85). Conversely, unlikely properties can be typical – just recall the shark example. The point is not that subjective typicality can deviate from real-world frequency (which is to be expected), but from subjective probability. A simple thought experiment makes this vivid. Imagine a subjective probability distribution over a one-dimensional space of death caps. Suppose the prototype is light green, and the density peaks there. Now add a second, slightly lower mode over a yellow region, reflecting the judgment that yellow death caps are also extremely likely. In this setup, yellow death caps are less typical than, say, dark-green ones (because farther from the prototype), yet they are judged more probable. Hence subjective typicality and subjective probability may diverge.

This shows that the authors' point is substantive. What is surprising, however, is the suggestion that the same point does not apply to expectation: In the thought experiment, I would also *expect* yellow death caps more than dark green ones, despite the latter being closer to the prototype.[21] The key issue is that expectation is conceptually tied to how likely one takes something to be. A model of expectation orderings that does not reflect this link is, at best, counterintuitive and, at worst, theoretically impoverished. As far as I can tell, Osta-Vélez and Gärdenfors (2022) do not argue for severing that link. Even if prototypes would coincide with points or regions of highest density – which is contentious[22] – the thought experiment shows that additional background knowledge about frequencies – perhaps modeled by a multimodal probability density function (see next section) – is sometimes more important for real-life reasoning than conceptual knowledge about typicality alone. Thus, I agree with the authors that expectations are structured on the basis of background knowledge, but such knowledge *need not be purely conceptual*.

None of this implies that PDP is a non-starter. Prototypes plausibly carry probabilistic information about member properties: "if the only thing we know about x is that it falls under concept M, we will expect it to be (close to) P^M, i.e., to have all the properties of the prototype." (Osta-Vélez & Gärdenfors, 2022, p. 84). Given the aims of this paper, I will not discuss any formal details here. Our analysis indicates that prototype-based priors are an idealization. As with PRS, they may aid normative theorizing, but their relevance to actual agents' reasoning is limited.

[21] I am grateful to Gottfried Vosgerau for raising this point in a conversation.

[22] Hampton (2006, p. 8) notes that "[…] that there were several different factors involved in determining mean typicality scores for common taxonomic categories like Bird or Fruit, including resemblance to other category members (as predicted by prototype theory) but also frequency of instantiation (how often the item is encountered) and fit to ideals (how well the item meets some goal or purpose – for example for artifact concepts)." It might also be instructive to distinguish stereotypes in lay concepts (sharks are dangerous) from prototypes in scientific concepts (sharks typically feed on mollusks and crabs), but that discussion is beyond this paper.

4.3 Analogical reasoning

Similarity is needed in theories of reasoning to account for certain inferences that people make (descriptively) or should make (normatively). A number of examples were already introduced, e.g., perceptual generalization in cognitive psychology. A long-standing topic in philosophy is analogical reasoning, which is a kind of inductive reasoning in which similarity relations play a key role. For instance, observing a member of *Amanita Phalloides* to be poisonous should lead me to expect not only that future conspecifics are poisonous, but also that *similar* Amanita species are poisonous. After arguing that two routes to priors (PRS and PDP) are descriptively limited, analogical reasoning will serve as a focal point for how similarity and probability might interact in real agents – with the caveat that the latter might merely approximate the introduced models. The key idea is to define and update probability distributions over hypotheses (or outcomes) on top of CS.

Once again, the account is inspired by Carnap: "Having rejected as hopeless earlier attempts to extend the system of λ-rules to incorporate analogical influence on the basis of syntax, Carnap (1980) recruited attribute spaces to model such influence semantically." (Douven et al., 2025, p. 185). To do so, he introduced two new rules to "derive the values of the rational confirmation functions for some simple propositions" (Sznajder, 2021, pp. 468–469) from the attribute spaces. The γ-rule states that prior probabilities should follow the PoI (Sect. 4.1). The η-rule captures analogy influence as depicted in the Amanita-example; roughly: Not only the predicate an observation falls under should benefit from the observation, but also its neighbors. How much a neighbor benefits depends on its distance (similarity) to the predicate in question, and is modeled by a probability distribution called the η-curve, which is a Gaussian distribution centered on the observed predicate.

Douven et al. (2025) extend this approach in three major ways. First, they utilize CS to allow more fine-grained observations. Carnap applies the distance measure between the centers of the *discrete* predicates which are the arguments of the η-curve. By contrast, Douven et al's analogy function takes points along *continuous* dimensions as arguments, and these can lie anywhere in the predicates. Second, while Carnap's η-curve is defined on a single dimension, Douven et al. accommodate any number of dimensions. And third, while Carnap considers only one observation, the authors cover iterative updating with multiple observations – resonating with tenet T3 of the PP framework, although the authors do not discuss this connection.

Rather than positing a hypothesis space (recall T&G's model), their model maintains a single probability distribution π_i over a conceptual space $\mathcal{S} \subseteq \mathbb{R}^n$, representing the agent's beliefs about where future observations are likely to fall. The prior π_0 is flat (geometric PoI). Let x be a sequence of point observations $\langle o^1, \ldots, o^n \rangle$ in the space, the update is:

$$\pi_i(x) = w\pi_{i-1}(x) + (1-w)d_i(x)$$

where $w \in [0,1]$ is a weighting parameter expressing inductive inertia (similar in spirit to Carnap's λ: the higher the w, the more weight is retained from past belief), and $d_i(x)$ is a Gaussian distribution centered at the new observation x_i. The dis-

tribution is modulated by a spread parameter that "leaves open the possibility that the analogy influence […] is different along the […] dimensions." (Douven et al., 2025, p. 12). Then, for any concept C representable in $\mathcal{S}$, the probability that the next observation $o^{i+1} \in C$ given x is calculated by integrating the current belief distribution over C:

$$\Pr\left(o^{i+1} \in C\right) = \int_C \pi_i(x)\, dx$$

To evaluate their model empirically, the authors test whether it predicts human responses in analogical reasoning tasks by reanalyzing experiments by Navarro et al. (2012). In these studies, participants were presented with data points that fall in a consequential region (e.g., bacterium A has survived temperature t) and asked to judge the likelihood that a second item also falls in that region (e.g., how likely will Bacterium B survive temperature t'?) – a setup that captures analogical effects. Douven et al. find a good correlation between their model's predictions and the empirical data, concluding that their geometrically grounded approach captures key patterns in human analogical reasoning. Despite its Carnapian lineage, their project is empirically oriented (C1).

While this approach is broadly Bayesian as it is an extension of inductive logic, it does not calculate posteriors via Bayes' rule. By contrast, Sznajder (2017, ch. 4) develops a „simple model" of inductive reasoning that is more explicitly Bayesian.[23] She defines a hypothesis space Θ consisting of unimodal beta distributions over a one-dimensional CS (the interval $[0, 1]$). The hypotheses are indexed by a continuous parameter $\theta \in \Theta$ giving the mode's location. Given observations $o^{1:n}$, the posterior is:

$$P\left(h_\theta | o^{1:n}\right) = \frac{P\left(o^{1:n} | h_\theta\right) P\left(h_\theta\right)}{\int_\Theta P\left(o^{1:n} | h_\theta\right) P\left(h_\theta\right) d\theta}$$

If the observations are points, their likelihood $P\left(o^{1:n} | h_\theta\right)$ is obtained by evaluating the beta density at each point. If observations are intervals (uncertain regions), the likelihood integrates the beta density over those intervals.

Note that Sznajder's model aims for a fully general account of sequential prediction on CS, while analogical reasoning is a special case that falls out if the hypotheses are Gaussian distributions. While Douven et al. use a single probability distribution over the space, Sznajder's model includes a space of hypotheses about what distribution is responsible for the observed data. Thus, it is more aligned with orthodox Bayesian approaches in cognitive science and arguably more compatible with stan-

[23] In her 2021 work, Sznajder generalizes her approach using Bayesian non-parametric techniques. She proposes using a Dirichlet process prior, an infinite-dimensional prior over distributions. A Dirichlet process can be thought of as a distribution on the space of all probability distributions on a continuum. This allows very flexible hypothesis spaces, and affects some of the math, but the outlines of her model remain intact, so I do not discuss this extension here.

dard formulations of PP, which typically rely on explicit inference over latent causes or generative models – see tenets T2 and T4.

Much more could be said in comparison of the two models, but I will focus on an aspect that relates to criteria of cognitive sparseness. While Douven et al.'s distribution represents a single best guess, Sznajder's is a "distribution of distributions" representing uncertainty over what the right guess should be. This layer exposes *predicate-hypotheses mismatches*: Carnap's self-similarity principle states that every predicate resembles itself the most. Consequently, a rule of analogy influence should be such that an observation that falls under P should confirm P the most. In Sznajder's model, irrational violations can occur if the posterior mode lies over a different predicate. She proposes moving predicate boundaries to the intersection points of competing hypotheses (where population density is lowest), yielding a mechanism for dynamic re-partitioning of the space. This aligns with Strößner's (2022) *mixed naturalness criterion*: not only geometric constraints (à la Gärdenfors) but also population density should guide conceptual carving – excluding, for instance, uninstantiated possibilities (e.g., twelve-legged horses) from shaping the natural concept (e.g., HORSE). Thus, while the additional hypothesis layer adds computational complexity, it comes with the advantage of accommodating Strößner's criterion.

Douven et al.'s framework, however, can be adjusted to similar effect: the minima of π_i mark low-density seams where boundary shifts are natural, so predicate mismatch could be handled by letting boundaries track minima. One may object that this is post hoc, whereas Sznajder builds in re-partitioning and grounds it by appeal to rational principles, such as self-similarity. On the other hand, Douven et al. also appeal to optimality as a driver of spatial partitioning; a parallel justification strategy may therefore be available. The normative-descriptive interplay resurfaces.

Both accounts illustrate how CS- and PP-style modeling can be partially integrated at a formal level, though Sznajder's is explicitly normative in aim. Empirical testing of Sznajder's model and deeper comparisons with the single-distribution approach are promising directions for future work.

5 Conclusion

Returning to the initial puzzle, here is how similarity and probability might interact in your mind. You might have a generative model of death cap mushrooms that consists of two components: A geometric space in which possible mushrooms are ordered by distance (similarity), and partitions of which represent properties or concepts. And second, a probability distribution over that space which encodes expectations, e.g., for what concept something falls under, given past experience. When you perceive something that fits your expectations – which assigns probability mass to your death cap region – you categorize it as a death cap and avoid it.

Empirically, aside from the reanalysis by Douven et al., we cannot yet say that agents in fact reason this way. On a theoretical level, however, the discussion demonstrates that CS and PP are partially methodologically compatible. A number of caveats must be highlighted, which are the reasons to speak of "partial" compatibility. The surveyed literature horizontally integrates CS with *elements* from PP – specifically,

Bayesian updating (T4), serving prediction-error minimization (T1) and active inference (T3). That said, Bayesian reasoning is not identical to PP. Notable differences include the roles of uncertain evidence and approximation strategies. Moreover, the surveyed models rely on rigid, non-hierarchical spaces, which is in tension with (T2). A natural direction for future work is to construct predictive models with hierarchical generative spaces and structural update mechanisms (e.g., for learning new dimensions). For a fuller integration of PP, formal aspects of the free energy formulation, which relies on variational Bayes as an approximation strategy, may also be relevant. For example, in Friston's models, the hidden causes an agent infers inhabit state spaces, which represent possible states of physical systems. The relationship between state spaces and CS remains to be explored, but lies beyond the goals of this paper. In light of the present discussion, a combined account which satisfies T1–T4 and C1–C3 is conceivable – CS can function as predictive models.

The issue of reductionism remains open. We considered what I take to be a convincing argument against reducing similarity to probability (pace T&G). I have not discussed work that takes similarity to be fundamental (e.g., Strößner, 2024, conference talk). Moreover, Poth's case turns on the explanatory value of cognitive content, which some may contest.

Another unsettled topic concerns Osta-Vélez and Gärdenfors' (2022) model of expectation orderings, which derives expectations from distance to prototype while ignoring probability and frequency information. If typicality strongly correlates with probability – which is plausible for property inheritance – disregarding probability may be misguided. Still, it is easy to envisage within-category distributions where the correlation breaks down. Open questions include whether PDP is an adaptive "fast-and-frugal" heuristic, or whether it works only under the idealization that no other information is available. It may help to treat typicality as a family-resemblance notion that sometimes aligns with conceptual information and at other times with probabilistic information.

Finally, I have only gestured at the relation between CS and Tversky's set-theoretic account of similarity, which matters for the paper's broader themes. In models that exploit distance to prototype, a clearer mapping would be useful. Strößner et al. (2021), for example, discuss inheritance between set-theoretic super- and sub-categories, which sits uneasily with (at least) perceptual CS, where concepts like *being green* do not have subcategories.[24] One option is to view "similarity as Bayesian inference [as an] opportunity to connect [Tversky's set-theoretic] feature-matching and geometric approaches by grounding them both in principles of statistical inference" (Poth, 2023c, Sec. 6). This paper has focused on connecting geometric similarity with PP; the connection looks promising, and further work is to be expected.

Acknowledgments I thank my supervisors and colleagues at the HHU, especially Gerhard Schurz and Gottfried Vosgerau, for their continued support and input. I would also like to thank Igor Douven, Matías Osta-Veléz and Peter Gärdenfors for valuable comments and discussions. Moreover, I would like to thank three anonymous reviewers and the editorial board members for a very productive review process which has improved the paper significantly.

[24] I am grateful to Gerhard Schurz for pointing this out in a conversation.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding was provided by my institution, the Heinrich-Heine-University Düsseldorf. No third-party funding was involved.

Data availability Not applicable.

Declarations

Competing interests I do not have any competing interests to declare.

References

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*, 6415. https://doi.org/10.1126/science.aat6766

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Brössel, P. (2017). Rational relations between perception and belief: The case of color. *Review of Philosophy and Psychology*, *8*(4), 721–741. https://doi.org/10.1007/s13164-017-0359-y

Bruineberg, J., & Rietveld, E. (Aug 2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00599

Carnap, R. (1952). *The continuum of inductive methods*. University of Chicago Press.

Carnap, R. (1967). *The logical structure of the world*. University of California Press.

Carnap, R. (1971). A basic system of inductive logic, part I. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability*. University of California Press.

Carnap, R. (1980). A basic system of inductive logic, part II. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability*. University of California Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/s0140525x12000477

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Colombo, M., Elkin, L., & Hartmann, S. (2021). Being realist about bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*, *72*(1), 185–220. https://doi.org/10.1093/bjps/axy059

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a grid-like code. *Science,352*(6292), 1464–68. https://doi.org/10.1126/science.aaf0941

Cubek, R., Ertel, W., & Palm, G. (2015). High-level learning from demonstration with conceptual spaces and subspace clustering. In *Proceedings - IEEE International Conference on Robotics and Automation* Vol. 2015. https://doi.org/10.1109/ICRA.2015.7139548

Decock, L., & Douven, I. (2011). Similarity after goodman. *Review of Philosophy and Psychology*, *2*(1), 61–75. https://doi.org/10.1007/s13164-010-0035-y

Decock, L., Douven, I., & Sznajder, M. (2016). A geometric principle of indifference. *Journal of Applied Logic*, *19*(2), 54–70. https://doi.org/10.1016/j.jal.2016.05.002

Douven, I. (2016). Vagueness, graded membership, and conceptual spaces. *Cognition*, *151*, 80–95. https://doi.org/10.1016/j.cognition.2016.03.007

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic*, *42*(1), 137–160. https://doi.org/10.1007/s10992-011-9216-0
Douven, I., & Gärdenfors, P. (2019). What are natural concepts? A design perspective. *Mind and Language*, (3), 313–34. https://doi.org/10.1111/mila.12240
Douven, I., Verheyen, S., Elqayam, S., Gärdenfors, P., & Osta-Vélez, M. (2025). Analogical reasoning: A carnapian approach. *Synthese*, *205*(5), 1–34. https://doi.org/10.1007/s11229-025-04991-y
Elqayam, S., Evans, J. S. B. T. (2011). Subtracting "Ought" from "Is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248. https://doi.org/10.1017/S0140525X1100001X
Facchini, A., & Termine, A. (2022). Towards a taxonomy for the opacity of AI systems. In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence 2021*. Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_7
Facchin, M. (2021). Predictive processing and anti-representationalism. *Synthese*, *199*(3), 11609–42. https://doi.org/10.1007/s11229-021-03304-3
Friedrich, J., & Fischer, M. (2024) Higher-level cognition under predictive processing: Structural representations, grounded cognition, and conceptual spaces. Preprint, 13 December 2024. https://doi.org/10.31219/osf.io/v436w
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622
Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11), e1000211. https://doi.org/10.1371/journal.pcbi.1000211
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787
Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, 20 YEARS OF fMRI, *62*(2), 1230–33. https://doi.org/10.1016/j.neuroimage.2011.10.004
Friston, K. (Nov 2002). Functional integration and inference in the brain. *Progress in Neurobiology*, *68*, 113–143. https://pubmed.ncbi.nlm.nih.gov/12450490/
Gładziejewski, P. (2021). Perceptual justification in the bayesian brain: A foundherentist account. *Synthese*, *199*(3), 11397–11421. https://doi.org/10.1007/s11229-021-03295-1
Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, *195*(6), 2627–48. https://doi.org/10.1007/s11229-016-1269-8
Gallagher, S., Hutto, D., & Hipólito, I. (2022). Predictive processing and some disillusions about illusions. *Review of Philosophy and Psychology, 13*(4), 999–1017. https://doi.org/10.1007/s13164-021-00588-9
Gauker, C. (2007). A critique of the similarity space theory of concepts. *Mind and Language*, *22*(4), 317–345. https://doi.org/10.1111/j.1468-0017.2007.00311.x
Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. *Simple heuristics that make us smart*. Oxford University Press.
Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.
Goodman, N. (1972). Seven strictures on similarity. *Problems and projects*. Bobbs-Merrill.
Gärdenfors, P. (1990). Induction, conceptual spaces and AI. *Philosophy of Science*, *57*(1), 78–95.
Gärdenfors, P. (2000). Conceptual Spaces: The Geometry of Thought. Bradford Books. Cambridge London.
Gärdenfors, P. (2011). *Inductive reasoning: From Carnap to cognitive science*. 1 January 2011.
Gärdenfors, P. & Osta-Vélez, M. (forthcoming). Reasoning with Concepts: Conceptual Spaces as a Framework. MIT Press.
Hampton, J. A. (2006). Concepts as prototypes. *46*, 79–113. https://www.staff.city.ac.uk/hampton/PDF%20files/Concepts%20as%20prototypes%202006.pdf
Hodson, R., Mehta, M., & Smith, R. (Feb 2024). The empirical status of predictive coding and active inference. *Neuroscience and Biobehavioral Reviews*, *157*, 105473. https://doi.org/10.1016/j.neubiorev.2023.105473
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–85. https://doi.org/10.1111/nous.12062
Jones, M., Bradley, C., & Love (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. https://doi.org/10.1017/s0140525x10003134
Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux. New York.
Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, *21*(2), 264–81. https://doi.org/10.1080/13869795.2018.1477983

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kriegeskorte, N., Rogier, A., & Kievit (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Margolis, E., & Laurence, S. (2021). Concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Spring 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/concepts/

Marr, D. (1982). *Vision: A computational investigation in the human representation of visual information*. Freeman.

Navarro, D. J., Matthew, J., Dry, & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223. https://doi.org/10.1111/j.1551-6709.2011.01212.x

Osta-Vélez, M., & Gärdenfors, P. (2022). Nonmonotonic reasoning, expectations orderings, and conceptual spaces. *Journal of Logic Language and Information*, *31*(1), 77–97. https://doi.org/10.1007/s10849-021-09347-6

Osta-Vélez, M., & Gärdenfors, P. (June 2020). Category-based induction in conceptual spaces. *Journal of Mathematical Psychology*, *96*, 102357. https://doi.org/10.1016/j.jmp.2020.102357

Poth, N. (2019). Conceptual spaces, generalisation probabilities and perceptual categorisation. In P. Gärdenfors, A. Hautamäki, F. Zenker, & M. Kaipainen (Eds.), *Conceptual spaces: Elaborations and applications*. Springer.

Poth, N. (2023a). Probabilistic learning and psychological similarity. *Entropy*, *25*(10), 10. https://doi.org/10.3390/e25101407

Poth, N. (2023b). Refining the Bayesian approach to unifying generalisation. *Review of Philosophy and Psychology*, *14*(3), 877–907. https://doi.org/10.1007/s13164-022-00613-5

Poth, N. (2025). Same but different: Providing a probabilistic foundation for the feature-matching approach to similarity and categorization. *Erkenntnis, 90*(1), 237–261. https://doi.org/10.1007/s10670-023-00696-1

Putnam, H. (1975). The meaning of "meaning". *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.

Quine, W. V. (1969). Natural kinds. In *Ontological relativity and other essays*. Columbia University Press. https://doi.org/10.7312/quin92204-006

Quine, W. V. (1986). Reply to Morton White. In L. Hahn & P. Schilpp (Eds.), *The Philosophy of W. V. Quine*. Open Court.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–41. https://doi.org/10.1073/pnas.0610341104

Robbins, H. E. (1992). An empirical bayes approach to statistics. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Foundations and basic theory*. Springer. https://doi.org/10.1007/978-1-4612-0919-5_26

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.

Scholz, S. (2024) Conceptual spaces: A solution to goodman's new riddle of induction? *Philosophia*. https://doi.org/10.1007/s11406-024-00744-2

Scholz, S. (n.d.). Conceptual spaces against semantic opaqueness in bayesian cognitive modelling. *Review of Philosophy and Psychology*.

Scholz, S., & Vosgerau, G. (2024). The cognitive and ontological dimensions of naturalness – Editor's introduction. *Philosophia*, *52*(4), 845–848. https://doi.org/10.1007/s11406-024-00775-9

Scholz, S., & Vosgerau, G. (2025). Conceptual spaces: Naturalness or cognitive sparseness? *Synthese*, *205*(3), 129. https://doi.org/10.1007/s11229-025-04941-8

Schurz, G. (2012). Prototypes and their composition from an evolutionary point of view. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality*. Oxford University Press.

Schurz, G., & Hertwig, R. (2019). Cognitive success: A consequentialist account of rationality in cognition. *Topics in Cognitive Science*, *11*(1), 7–36. https://doi.org/10.1111/tops.12410

Shapiro, L., & Spaulding, S. (2024). Embodied cognition. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2024/entries/embodied-cognition/

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–40. https://doi.org/10.1007/BF02289630

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219–46. https://doi.org/10.1007/BF02289621
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–23. https://doi.org/10.1126/science.3629243
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin (US)*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3
Sprevak, M., & Smith, R. (2023). An introduction to predictive processing models of perception and decision-making. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12704
Strößner, C. (2022). Criteria of naturalness in conceptual spaces. *Synthese*, *200*(78), 36.
Strößner, C. (2023). Hybrid frameworks of reasoning: Normative-descriptive interplay. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). https://escholarship.org/uc/item/58j925zr
Strößner, C. (2024) Similarity first. Conference talk. Conceptual Spaces at Work, HHU Düsseldorf. https://www.philosophie.hhu.de/fileadmin/redaktion/Fakultaeten/Philosophische_Fakultaet/Philosophie/Philosophie_Seniorprofessur_Schurz/Workshop_Conceptual_Spaces/Corina_Stroessner.mp4
Strößner, C., Schuster, A., & Schurz, G. (2021). Modification and default inheritance. In S. Löbner, T. Gamerschlag, T. Kalenscher, M. Schrenk, & H. Zeevat (Eds.), *Concepts, frames and cascades in semantics, cognition and ontology*. Springer International Publishing. https://doi.org/10.1007/978-3-030-50200-3_14
Sznajder, M. (2017). *Inductive logic on conceptual spaces*. Rijksuniversiteit Groningen.
Sznajder, M. (2021). Inductive reasoning with multi-dimensional concepts. *British Journal for the Philosophy of Science*, *72*(2), 465–484. https://doi.org/10.1093/bjps/axz012
Tenenbaum, J. B., Thomas, L., & Griffiths (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/s0140525x01000061
Tětková, L., Brüsch, T., Scheidt, T. K., et al. (2025). On convex decision regions in deep network representations. *Nature Communications*, *16*(1), 5419. https://doi.org/10.1038/s41467-025-60809-y
Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–52. https://doi.org/10.1037/0033-295X.84.4.327
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124
Wason, P. C. (1969). Regression in reasoning? *British Journal of Psychology*, *60*(4), 471–480. https://doi.org/10.1111/j.2044-8295.1969.tb01221.x
Yellapragada, S., & Prakash, K. C. (2019). Variational bayes: A report on approaches and applications. Preprint. https://www.semanticscholar.org/paper/Variational-Bayes%3A-A-report-on-approaches-and-Yellapragada-Prakash/5e034ac4ec062fea0c37587724de0da713a9eb0d

Part II – Publications
Paper D

Semantically Deficient Models of Cognition – A Bayesian Case Study, Conceptual Spaces, and AI Opacity

Bibliographic reference. Scholz, S. (2026b). Semantically Deficient Models of Cognition – A Bayesian Case Study, Conceptual Spaces, and AI Opacity. Review of Philosophy and Psychology. Advance online publication.
DOI. https://doi.org/10.1007/s13164-026-00834-y
Authorship. Sole-authored by Sebastian Scholz.

Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-026-00834-y

Semantically Deficient Models of Cognition – A Bayesian Case Study, Conceptual Spaces, and AI Opacity

Sebastian Scholz[1]

Received: 7 November 2025 / Accepted: 19 June 2026

Abstract
Nina Poth criticizes Tenenbaum & Griffiths' Bayesian model of perceptual generalization for its "semantic opaqueness" – roughly, it is unclear what the model's probabilistic states are about – which suggests a connection with AI opacity. The first half of this paper analyzes the connection and suggests "semantically deficient models of cognition" (SEDMOC) as improved terminology. Crucially, for some purposes and in some contexts, models of cognition should reveal to cognitive scientists how semantic content produces behavior. While this is not at issue in the standard opacity debate, semantic opacity is one potential source of deficiency. The second half of this paper investigates hybrid models that combine Bayesian reasoning with geometric similarity (more specifically with Conceptual Spaces), thereby avoiding deficiency, and proposes to let the probability function shape spaces in a hierarchical model, outlining a PCA-based abstraction mechanism. It is argued that the proposed mechanism may introduce opaque dimensions which, however, are not semantically deficient. Both the terminological shift and the PCA-based abstraction mechanism contribute to better model-based explanations of cognitive behavior.

1 Introduction

Bayesian models of cognition use probability theory to explain cognitive capacities and articulate norms of rational inference. They are applied both widely and successfully. A forceful criticism, however, is that they risk devolving into *just-so* stories, mathematical frameworks so flexible that they can be made to fit any data (cf. Bowers & Davis 2012, Crupi & Calzavarini 2023, Hahn 2014; Jones & Love 2011).

Poth (2019, 2023) formulates a version of this critique that is particularly noteworthy because her terminology suggests a connection with recent debates over AI-

✉ Sebastian Scholz
s.scholz@hhu.de

[1] Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Published online: 07 August 2026

opacity. Her primary target is Tenenbaum & Griffiths' Bayesian model of perceptual generalization (2001, henceforth T&G). T&G posit a set of hypotheses ordered by the relative "sizes" of candidate categories, yet they leave open "how the size of a candidate category should be measured, and its semantic content interpreted" (Poth 2019, p. 10). As a result, it is unclear what the model's probabilistic states – or other formal elements – are about. This is roughly what Poth calls "semantic opaqueness".[1]

The first part of this paper (Sect. 2) analyzes Poth's critique and its connection with AI opacity. It suggests "semantically deficient models of cognition" (SEDMOC) as improved terminology. The insight here is that, for some purposes and in some contexts, models of cognition should reveal to cognitive scientists how semantic content produces behavior. While this is not at issue in the standard opacity debate, semantic opacity is one potential source of deficiency.

The second part of the paper is constructive. It *picks up on* and *applies* Poth's proposed solution of semantic opaqueness by (i) examining in Sect. 3 models from the literature on Conceptual Spaces (CS) – a similarity-based account of conceptual content – that integrate similarity with probabilistic inference, and by (ii) outlining in Sect. 4 a semantically interpretable model of abstraction based on Principal Component Analysis (cf. Gärdenfors & Osta-Vélez, submitted, 2026). The model is interesting because it addresses a research gap: While other models keep probability and similarity relatively separate, this one allows the probability function to shape spaces in a hierarchical model, outlining a PCA-based abstraction mechanism – without collapsing the distinction between the two types of information. The reason to discuss it *here* is that the proposed abstraction mechanism might introduce new dimensions that are difficult to interpret. I argue that this does not lead to semantic deficiency as defined in the first part of the paper. Both the terminological shift and the PCA-based abstraction mechanism contribute to better model-based explanations of cognitive behavior.

2 The Nature of the Problem

This section presents T&G's model, analyzes Poth's critique, works out its connection with AI opacity, and advocates for the philosophy of explanations in cognitive science which underlies it. It concludes that Poth's critique makes sense, but should run under a different label.

2.1 Case Study and Critique

T&G engage Shepard's (1987) universal law of generalization: The likelihood that an organism generalizes a type of behavior from one stimulus to another drops exponentially with the dissimilarity between them. This generalization gradient is universal in the sense that it shows up across a diverse range of stimuli and even species. Shepard explains it via psychological similarity modeled as spatial distance.

[1] Throughout the paper, I will use "opaqueness" to signify Poth's critique, and "opacity" to refer to the notion that AI systems are black boxes (see Sect. 2.2).

For example, suppose you encounter a poisonous mushroom and learn to avoid it. Should you generalize avoidance to a second mushroom? In such cases, one cannot know which mushrooms *truly* must be avoided. Instead, one must "infer the conditional probability that an unknown instance, y, falls under a candidate concept, C, given that a previously observed example, x, falls under C" (Poth 2023, p. 3). In philosophical terms, one must approximate the *intension* of the concept since the true *extension* is epistemically inaccessible.[2] Shepard's proposal is that this is accomplished via "an internal metric of similarity between possible situations" (Shepard 1987, p. 1317). Presented with the second stimulus, one consults this metric – as one would use a map. If the spatial distance between stimuli on the map is small, they are similar, generalization is likely, and C can be treated as a region of unknown boundaries. This "map" resembles a conceptual space, but the proponents of CS tend to take geometric similarity seriously as an organizing principle of the mind, while Shepard (1987) viewed it more as a modeling tool (see Sect. 3.1).[3]

In contrast, T&G aim to explain the gradient without invoking a specific conception of similarity. They construe it as a subjective probability distribution that learners maintain over a space of hypotheses about what the true C is, updated in the light of evidence – a paradigmatic Bayesian approach. Bayes' theorem states that the posterior probability of a hypothesis given evidence $Pr\,(h|x)$ is proportional to the product of its prior probability without evidence $Pr\,(h)$ and the likelihood $Pr\,(x|h)$, which is the probability of the evidence given the hypothesis:

$$Pr\,(h|x) = Pr\,(x|h) \times Pr\,(h)$$

(For present purposes, we set the normalizing denominator $Pr\,(x)$ aside). T&G use a modified version of the theorem, replacing the likelihood by the *size principle*. They compute the generalization function:

$$Pr\,(y \in C|x) = \sum_{h:y \in h} Pr\,(h|\,x)$$

which yields the exponentially decreasing curve. The summands are "the probabilities $p\,(h|x)$ of all hypothesized consequential regions that contain y" (T&G, p. 631). They are posteriors which are calculated with the modified Bayes-rule: For n observations $X = \{x_1, \ldots, x_n\}$, T&G set the likelihood.

[2] Concepts are the building blocks of thought (cf. Margolis & Laurence, 2021) which cognitive psychology ascribes to explain *higher* cognitive capacities such as categorization and reasoning. *Perceptual* generalization may not count as a higher capacity, but this distinction is not very important here. To keep things simple, I will continue to speak of perceptual concepts.

[3] Shepard's thinking on this point, however, has developed (cf. 1981, 1994). Moreover, (Gärdenfors, 2000) explicitly embraces instrumentalism, while later work suggests an 'organizing-principle outlook' – as is discussed below. Tracing and comparing these developments is a promising project beyond this paper. I am grateful to Nina Poth for highlighting this point in a conversation.

The idea behind this formula is that "smaller hypotheses receive higher likelihoods than larger hypotheses, by a factor that increases exponentially with the number of examples observed" (T&G, p. 634). That the number of examples matters should be intuitive. The insight of the size principle is that a small region that contains some set of observations is more likely to produce that set under random sampling than a large region that contains the same set. In other words, the size principle concerns the information value of the hypotheses, and tells us why it is rational for learners to prefer smaller hypotheses. To make this concrete:

> […] all else being equal, it seems less probable that a praying mantis will fly into my visual field than that any random insect will. Hence, $Pr\,(h_P) < Pr\,(h_I)$. However, were a praying mantis to fly into my visual field now, this observation would confer a higher degree of confirmation on h_P than on h_I (even though both hypotheses are compatible with the observation), and so $\Pr\,(e|h_P) > \Pr\,(e|\,h_I)$. (Poth 2023, p. 4).

T&G emphasize that their account subsumes different conceptions of similarity: it purportedly does not matter "based on which quantity the size of a consequential region is actually measured" (Poth 2019, p. 14). Poth disagrees by highlighting the "semantic value" of similarity "for interpreting the size principle" (p. 16). On T&G's account, when the model "observes" a praying mantis, the observation is merely an element of a set that drives updating in accordance with probabilistic norms and the size principle. The model does not say why this element belongs to one set rather than another, why one set is smaller than the set of all insects, or, more generally, what the model's probabilistic belief states are about beyond numerical values.

When *we* observe a praying mantis, by contrast, we treat it as instantiating a concept because of its similarity to *meaningful* experiences. Indeed, the similarity relations between experiences might be what constitutes the concept in the first place. If conceived geometrically, they provide a precise and transparent measure of region size and degree of confirmation. Poth (2019) proposes a modified version of the size principle that "defines the content of the hypotheses more precisely: as mappings from measurable regions in conceptual space […] to the labels whose meaning is to be inferred […]" (p. 21). Its upshot is to measure the relative spatial overlap between evidence and target concept, which "helps to make more explicit the relationship between the evidence and a hypothesised category […]" (ibid.), thereby increasing semantic transparency. Region-size itself "is measured by integrating over the subset of vectors in conceptual space that would be covered by each category predicate" (p. 7). Section 3 discusses this solution in more detail.

One might reply that T&G simply measure region size by counting elements, obviating geometric similarity. However, counting elements does not work if concepts are thought to contain continuum possible elements. Moreover, the point is that similarity is doing the work of determining into which sets the elements belong. This should be made explicit rather than tacitly assumed. As Jones and Love (2011) put it, "Bayesian inference itself is conceptually trivial, but strong assumptions are often embedded in the hypothesis sets and the approximation algorithms used to derive model predictions, without a clear delineation between psychological commitments

and implementational details" (p. 169). In T&G's case, the missing psychological commitments concern the role of similarity and the efficacy of cognitive content, while Jones and Love emphasize mechanistic considerations. The broader lesson is that semantic opaqueness is a special case of a more general concern about Bayesian cognitive models. It applies beyond T&G – Poth (2023) discusses further cases – but T&G serve here as a suitable paradigm case.

It is noteworthy that the 2023 paper does not use the term opaqueness, but it links the same points discussed under this label by the 2019 paper to "good old-fashioned" AI (GOFAI). GOFAI systems manipulate symbols according to syntactic rules. Fodor (1975) famously views the mind as a GOFAI system, arguing that concepts are innate expressions in a language of thought. Tenenbaum and Griffiths (T&G) can be situated within the symbolic paradigm insofar as their model manipulates belief states according to the syntactic rules of probability theory (cf. Poth 2023, Sect. 2.1).[4] Crucially, symbol manipulation works regardless of what the symbols mean. It is thus unclear what explanatory role the semantics plays in such models, and where it ultimately comes from. This is the *symbol grounding problem*:

> How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? (Harnad 1990, p. 335).

I thus contend that semantic opaqueness means two distinct things which are intertwined in Poth's writings:

Semantic opaqueness O1: *By not explicitly defining a semantics for their probabilistic model, T&G obscure what informs the structure of their hypothesis space.* The problem here is not the symbolic nature of the model. Instead, the problem is that the account specifies *that* some concept C has a smaller size than some other candidate concept, and this is the condition for the generalization rule to determine that the (rational) agent must prefer a hypothesis stating that concept. But the model lacks an accompanying account of *why* that concept has a smaller size than the alternatives. Because the model does not specify this, it doesn't give the answer to a relevant why question that concerns the concept's semantic content, the learner's learning history, and the optimality constraints (see Sect. 2.3 on the significance of why questions).[5] O1 can be fixed by transparently defining a semantics.

Semantic opaqueness O2: *Because any semantics would be extrinsic to T&G's model, it runs into the symbol grounding problem.* This problem cannot be fixed by transparently defining a semantics for the model.

At first glance, these two points are very different, and one may wonder why Poth connects them at all. Section 2.3 argues that doing so is the right move, but let us first explore the relationship between semantic opaqueness and AI opacity.

[4] It should be pointed out that T&G operate within the Computational-Representational Theory of Mind, which does not necessitate symbolic computation.

[5] I am grateful to an anonymous reviewer for this formulation.

2.2 AI Opacity

Opacity means that the relevant systems are "black boxes" in various ways, some of which will be specified below. As a result, it is difficult to understand how the outputs of the systems come about, and whether the systems are aligned with our own goals – raising various practical and ethical concerns. The antithesis to opacity is interpretable, or explainable AI (XAI).

In the XAI literature, GOFAI models are generally regarded interpretable. Both the semantics of the symbols, which are standardly defined by the modeler, as well as the calculation steps leading to outputs, are traceable. What the literature doubts instead is the interpretability of *connectionist* systems. Connectionism "is a movement in cognitive science that hopes to explain intellectual abilities using artificial neural networks […], simplified models of the brain composed of large numbers of units (the analogs of neurons) together with weights that measure the strength of connections between the units." (Buckner & Garson 2025, p. 1). In contrast to classical architectures, they do not have atomic symbols which can carry semantic information, but the information is distributed throughout the acquired weights and thresholds in trained neural networks. As a result, they do not suffer from the symbol grounding problem, but the distributed nature of the information, as well as the sheer complexity of contemporary systems, have led to the vast and terminally confused debate at hand (Lipton 2018).

At this point, one may conclude that semantic opaqueness has nothing to do with AI opacity – despite superficial terminological similarity – because symbol grounding is not at stake (O2), and it doesn't make sense to ask for explicit definitions of semantics for non-symbolic systems (O1). But that would be too hasty. Consider the typology of opacities by Facchini and Termine (2022). A first important distinction concerns knowledge *about* models versus knowledge *with* models. "Access opacity" corresponds to the former and concerns the "capability of understanding the structure and functioning of an AI system" (p. 5). It may be caused by restricted access to training data, a lack of background knowledge or skill on part of the stakeholder, or the complexity of the model. Access opacity is what most XAI accounts deal with.[6] "Link opacity" concerns the link between model and target. Many AI models are mere prediction engines that do not tell us on a deep level how and why phenomena occur, impeding knowledge *with* models about their scientific targets (such as targets in physics or medicine, e.g., Sullivan 2022). Finally, "semantic opacity" "happens […] if the format used to store and manipulate information prevents users from giving it a meaningful interpretation […]" (Facchini and Termine 2022, p. 12) – either because there is no well-defined semantics, or the users lack the skills to understand it, or it is inadequate for the context.

It should be clear that T&G's model is not access-opaque to cognitive scientists. And it is also true that researchers who discuss link opacity mostly care about connectionist systems. However, O1 *does* resonate with the characterization of semantic

[6] "However, the meaning they attribute to the term „opacity" is clearly different from the usual one in XAI. It does not refer to users' understanding of the ML model but to users' (in)ability to understand something about target-phenomena with a ML model." (Facchini & Termine, 2022, p. 4).

opacity. Connectionist models can frustrate semantic interpretation because contents are epistemically hard to access when distributed across a network. T&G's model, by contrast, is *meaningless* in the sense that "there is no well-defined semantics", even though there could be, which does prevent users from giving the priors, probabilistic states, etc. meaningful interpretations. Strikingly, the users here are cognitive scientists who target cognitive behavior. This is a special case rarely discussed in the XAI literature, which is more concerned with practical or ethical issues such as alignment. In this special case, I shall argue, semantic opacity *constitutes* a form of link opacity, although Facchini & Termine treat the "three macro-forms of opacity" as conceptually and logically independent (p. 14). The reason is that explanatorily valuable models *of* cognition (as opposed to cognitive models in the broad sense of smart artifacts) need interpretable cognitive content to explain their targets. This is also the thread that connects O1, O2, and semantic opacity. The next section elaborates.

2.3 Semantic Deficiency Undermines Explanatory Value

T&G's model is hard to interpret – but people are, too. According to the received view, we do not have direct access to other minds, and we therefore postulate mental states such as beliefs and desires about beer to explain why somebody went to the fridge, for instance (cf. Hutto & Ravenscroft 2021). In cognitive psychology, it is standard to postulate mental representations to explain flexible behavior (Newen & Vosgerau 2020). The point is that both everyday and scientific explanations appeal to states with cognitive content. *Cognitive* means we are talking about something that occurs within cognitive systems (roughly: systems that show problem-solving behavior). And "*[c]ontent* is a deliberately vague term; it is a rough synonym of another vague term, 'meaning'. A state with content is a state that *represents* some part or aspect of the world; its content is the way it represents the world as being." (Brown 2022, Sect. 1). Thus, representation implies the possibility of misrepresentation, but its conditions don't have to be absolute; they can be indexed to pragmatic aims in navigating a resistant world (cf. Williams 2018).

One might doubt whether positing such representations is warranted. I do not attempt a full defense here, but sketch the underlying rationale. Natural phenomena typically admit explanations at multiple levels. The color of a flower depends on microstructural properties of pigments and on selective pressures from pollinator vision. As Sober (1999) emphasizes, adequate scientific understanding often requires both high-level and low-level explanations, where the relevant level depends on pragmatic research interests (Newen & Vosgerau 2020, p. 192). Cognitive psychology frequently trades in high-level explanations that invoke content. The intelligibility of such explanations depends on the flexibility of the behavior at issue. Frogs snap at any black moving dot; such rigid behavior can be captured by relatively simple physiological accounts. By contrast, ants initiate special search routines when triangle completion fails (e.g., after displacement by an experimenter). Thus, "describing the ant's behavior as a homing behavior brings an explanatory advantage: it allows us to systematically distinguish between successful cases and unsuccessful cases […]" (Newen & Vosgerau

2020, p. 197). In other words: This behavior requires high-level explanations that involve the ascription of cognitive content, i.e., nest representations. They cannot be reduced to low-level explanations because the high-level states are multiply realizable and thus cannot be individuated at lower levels.

Interestingly, Egan (2014) argues that cognitive content is an unscientific gloss, and only explanations appealing to mathematical content are scientific – a consideration that could vindicate T&G's model, given that it has mathematical but no cognitive content. Newen and Vosgerau (2020) highlight a number of problems with this view, a general one of which shall be restated here:

> [I]f we describe a physical mechanism with a mathematical operation, then the latter remains underdetermined relative to the mechanism, since usually there are many mathematical functions which fit the same data of the mechanism. Our selection of one mathematical operation for an explanation depends on other presuppositions, including the selection of an explanatory interest. (P. 191).

This mirrors the multiple-realization argument: mechanisms are multiply "mathematizable." It already suggests why a model with only mathematical content is problematic in cognitive psychology: it fails to deliver the irreducible, high-level explanatory resources the field often requires. The point becomes clearer by reflecting briefly on *models* and *explanations*.

Scientific models are heterogeneous, but they are often treated as idealized representations that enable surrogative reasoning about target systems (Frigg & Nguyen 2021). They may mediate between theories and data (Morrison & Morgan 1999), and on some views theories are collections of models.[7] Crucially, models are standardly taken to serve epistemic aims such as explaining or understanding their targets (Frigg & Hartmann 2025; Sect. 3.2–3.4). As per usual, these notions are contested; here I focus on explanation. Schurz (2014) conceives explanations as answers to why-questions that, prototypically, render explananda expectable, specify causes, and unify phenomena – goals that can conflict. Pragmatic theories add that evaluative standards vary with context:

> Let us say that a theory of explanation contains "pragmatic" elements if (i) according to the theory, those elements require irreducible reference to facts about the interests, beliefs or other features of the psychology of those providing or receiving the explanation and/or (ii) irreducible reference to the "context" in which the explanation occurs. (Woodward & Ross 2021, Sect. 6).

A simple (fictional but realistic) example illustrates these ideas.

Dr. Alvarez trains a model to predict mushroom toxicity from physical features: cap color, shape, presence of gills, etc. Initially, predictions are unreliable. After trying various options and setting a parameter to $p = 2.2$, performance becomes reliable.

[7] This is the "so-called semantic view of theories that came to prominence in the second half of the 20th century" (Frigg & Nguyen, 2021; Sect. 4).

I. Why does the model now predict mushroom toxicity?

a) Because parameter p was set to 2.2
b) Because the *learning rate* was set to 2.2
c) Because it captures correlations between mushroom features

I. is a why-question about the model, and Ia. – Ic. are candidate explanations. Suppose that none of the explanations are false. Yet, Ic. is more relevant to a biologist than Ia., while it might be less relevant for a programmer. Interestingly, Ib. provides a semantic interpretation for p, which adds explanatory depth – perhaps because it connects p with further knowledge (e.g., about how deep neural networks work) in the service of unification.

II. Why is mushroom m toxic?

a) Because of surface features F
b) Due to selective pressure from snails

III. Why are mushrooms with surface features F toxic?

Here the explanandum is the target system. Predictive success alone does not guarantee that a model reveals the causal structure of its target (recall *link opacity*). If Dr. Alvarez's system is a mere prediction engine – as many popular AI models are – it can make explananda expectable and thus count as explanatory in Schurz's prototypical account. That may suffice for IIa. under a statistical framing. However, IIb. is what an evolutionary biologist seeks, and the model as such does not deliver it. Similarly, the second "why" question III. calls for deep causal explanation.[8] The example thus fixes a context (e.g., evolutionary biology) in which mere prediction is insufficient. This supports clause (ii) of Woodward & Ross's characterization, although it might be debated whether reference to the context is strictly irreducible here.

Now consider a cognition-targeted question:

IV. Why did cognitive system S categorize mushrooms with features F as toxic?

When the explanandum is a cognitive capacity, the research context (cognitive psychology or a neighboring field) typically requires explanations of flexible behavior that appeal to representational contents, as argued by Newen and Vosgerau (2020). Models that support such explanations should therefore *reveal to cognitive scientists how content is efficacious in producing behavior*. Call this the explanatory constraint (EC). Semantically opaque prediction engines do not satisfy EC, whether

[8] As an instructive historical example, consider *Newton's laws*. While enabling prediction of the movements of celestial bodies, thereby satisfying traditional accounts of explanation (e.g., the *deductive-nomological* account holds that predictions must logically follow from laws and initial states, see Woodward & Ross 2021), many of Newton's contemporaries considered the laws explanatorily deficient because they do not specify mechanisms by which gravitational forces, for instance, take effect (Pernu, unpublished manuscript).

they are artificial neural networks or Bayesian machinery (the latter of course corresponds to O1). If people are, for practical purposes, black boxes and we build cognitive models to peer inside, those models should not be black boxes themselves. At minimum, given equal predictive performance, we should prefer the more interpretable model.

Although O2, in contrast to O1, is not a matter of interpretability, it also frustrates EC: Because model performance depends only on syntactic calculations, the semantics is both causally and explanatorily impotent due to the good old-fashioned nature of the model. The model's internal operations make no reference to what the symbols represent, the content is detached from the defining mathematics. Of course, some actors in the philosophy of mind embrace this *epiphenomenalism*, e.g., "Stich proposes a *syntactic* theory of the mind, on which the semantic properties of mental states play *no* explanatory role" (Pitt 2020; Sect. 2). Without entering the metaphysics of mental causation, the foregoing considerations show why such view is highly unsatisfactory. Thus, given equal predictive performance, we should prefer the non-symbolic model.

The discussion suggests three major insights. First, Poth's concept of semantic opaqueness is useful because it sheds light on what we demand of models which aid explanations of cognitive behavior – GOFAI, Bayesian, and connectionist models. However, it became clear that this is at best a small subset of what is discussed under AI opacity, namely, the subset where cognitive scientists develop models *of* cognition. Thus, the term "semantic opaqueness" can be misleading. I hereby suggest to speak of "semantically deficient models of cognition" instead (SEDMOC), which is more specific and gets rid of false associations such as access opacity and the narrow focus on connectionist systems.

Second, SEDMOC should be considered link opaque, however. Link opacity occurs when the lack of interpretability of a model impairs understanding of a scientific target system (e.g., the weather system in opaque climate models). Qua definition, SEDMOC target cognitive systems. I have argued that such models must satisfy EC to enable us understand these targets. Semantically opaque models in the sense of Facchini and Termine (2022) do not satisfy EC. Thus, in the special case when cognitive behavior is the target and cognitive scientists are the stakeholders, semantic opacity constitutes a form of link opacity.

Third, the critique of SEDMOC sits well with a pragmatic philosophy of explanation – not only because EC is addressed to a particular research context, but also because the demand for semantic interpretability is audience-relative, tracking the psychological features of stakeholders – clause (i) in Woodward & Ross's characterization. The standard cannot be framed from a stance-independent point of view: *interpretable or deficient for whom* and *for what purposes*? Endorsing pragmatism does not collapse standards into subjectivism; it simply acknowledges that what counts as a good explanation is partly sensitive to context and audience. While I have not discussed whether proponents of the larger just-so story critique of Bayesian cognitive science would endorse a pragmatic philosophy of explanation, I do think that to the degree something like EC drives their critique, it would suit them well.

3 Towards a Solution

The paper so far has analyzed and endorsed the critique of semantic opaqueness under a refined label. What follows, first, introduces models which are not semantically deficient and discusses why this is so, and second, applies insights from that discussion by introducing a new model of abstraction under probabilistic CS, and evaluating its compliance with EC.

3.1 Enter Conceptual Spaces

Poth's proposed remedy is to augment the probabilistic model with "a model of the representational content associated with concepts (i.e., what was yet missing from probabilistic models when used for the purpose of explaining psychological behavior)" (Poth 2023, p. 8), and the framework she advocates are CS (Gärdenfors 1990, 2000, 2014). Representational content in CS is structured geometrically. The similarity of objects is represented by their spatial proximity along so-called *quality dimensions* which represent the various qualities that objects can have. Concepts are identified with *regions* in the spaces. For example, the concept GREEN is a region in the color space, consisting of three dimensions: hue, saturation, and brightness. The framework is mathematically precise, accommodates prototypes (roughly, best examples of categories), and supplies a theory of conceptual naturalness or "cognitive sparseness" (Scholz & Vosgerau 2025). Natural concepts are necessarily *convex*: if two points lie in a region, every point on the path between them also lies in that region. Such convex partitions are argued to arise from principles of cognitive economy – balancing informativeness, parsimony, and related goals under resource constraints (Douven & Gärdenfors 2019).

CS is not a mathematical just-so story because it carries explicit psychological commitments beyond bare geometry. For example, it yields falsifiable predictions on how cognitive content (e.g., the shape of the GREEN region) informs inductive reasoning about green objects. A productive way to frame the point is to locate T&G on the *computational* level of Marr's (1982) explanatory analysis, and CS on the *algorithmic* level. In Marr's scheme, the computational level specifies *what* is being computed and *why.* The algorithmic level specifies *how* the problem is solved – the representations used and the processes that transform them. Higher levels are thought to be multiply realizable – but constrained – by lower ones. Bayesian models often leave representational format and psychological content unspecified, yielding semantic opaqueness through their computational focus. Recall that T&G regard it as a feature of their account not to commit to a specific similarity-based representational format. By contrast, CS is best understood at the algorithmic level because it specifies semantically interpretable representational primitives (quality dimensions, metrics, etc.) and admissible procedures over them for tasks such as categorization and induction.

Accordingly, adopting CS would commit T&G to psychological entities structured by similarity – structure which has been implicitly assumed – thereby avoiding O1. Note that this solution does not work if geometry is *only* used as a modeling tool. Presumably, that's why Poth (2023, Sect. 2.1) includes Shepard (1987) as a target of

her criticism: Geometry qua mathematical tool doesn't provide the theory of concept learning that's needed to explain generalization behavior.

Adopting CS, however, also provides a plausible response to O2. Bermúdez (2014) outlines two basic strategies: "we can either abandon the idea that cognition is a form of information processing (and with it abandon the idea that cognitive science can explain the mind). Or we can look for forms of information processing that are not symbolic" (p. 167). The locus classicus of the latter is connectionism:

> One of the attractions of distributed representations in connectionist models is that they suggest a solution to the problem of providing a theory of how brain states could have meaning. The idea is that the similarities and differences between activation patterns along different dimensions of neural activity record semantical information. So the similarity properties of neural activations provide intrinsic properties that determine meaning. (Buckner & Garson 2025; Sect. 8).

However, activation patterns belong to the *implementational* level (concerning the physical realization of algorithms), one level of explanatory analysis below the CS account. Moreover, connectionism undermines EC: contents are distributed and therefore hard to localize, and their efficacy in producing behavior is less transparent. Gärdenfors positions CS *between* symbolic models and neural nets: symbols map onto CS, and CS abstractions arise from high-dimensional neural activation patterns. Crucially, like activation patterns, CS regions are not arbitrary vehicles of externally imposed content, but are *intrinsically meaningful* in the sense that: "the representing relation has the same inherent constraints as its represented relation" (Gärdenfors 2000, p. 44).[9]

Even with intrinsic representation as a technical response to O2, however, a potential hybrid model would not automatically satisfy the explanatory constraint EC. To see this, consider Goodman's (1983) concept GRUE, defined as: "applies to all things examined before t just in case they are green but to other things just in case they are blue" (p. 74). The region representing GRUE is intrinsically meaningful, yet arbitrary in the sense of gerrymandered. A cognitively adequate model of reasoning would have to rule this out, leaving a *sparse* subset of all possible concepts (see Scholz & Vosgerau 2025), but nothing in the model at this point allows it to do so. Exclusion would again have to be achieved externally – by further semantic assumptions that enter the model in a potentially opaque way.

[9] I do not assume that CS provide a complete solution to the symbol grounding problem, nor that they settle general questions about intentionality or reference. The point is rather that, for perceptual domains, CS representations are not arbitrary symbols whose meanings are fixed only by external interpretation. Their dimensions, metric structure, and regions are constrained by perceptual discriminability and similarity relations. This is the sense in which Gärdenfors suggests that conceptual spaces help with the grounding of perceptual symbols: the representing structure preserves relevant constraints of the represented domain. One may doubt whether this suffices for full semantic grounding, but the present argument requires only the weaker claim that CS provides a non-arbitrary and semantically interpretable representational format for perceptual contents.

The key to addressing this concern within CS is *optimality*. In contrast to pure formal rationality, optimality considerations bring in assumptions about the environment and the goals of the cognitive system. A model that is optimal doesn't just reason correctly in the abstract; it also uses the right priors, hypotheses, or representations so that its inferences succeed in the actual world. In cognitive science, John Anderson's notion of *rational analysis* is emblematic of this approach (see Chater & Oaksford 2000).[10]

Poth (2023) connects this with *directional structural resemblance* (or "second-order homomorphism"), the idea that internal similarity relations correspond to worldly similarities.[11] On her view, "structural similarity ensures a rational connection to the outside world" (p. 12) and enables effective action. Gärdenfors does not endorse such correspondence in a robustly realist sense:

> Representations in conceptual space are purely epistemic entities; the theory makes no commitment to the assumption that concepts represent the world in a truth-conducive manner. In contrast, structural-resemblance approaches additionally consider how internal geometrically structured representations must be arranged to fit the causal structure of the world and guide successful action. Poth (2023, p. 13).

Even so, Douven and Gärdenfors (2019) link CS to cognitive economy in a way that resonates with rational analysis. Their optimality criteria secure a rational connection not to the *distal* world in itself, but to a *proximal* environment – an agent-shaped reconstruction guided by pragmatic interests. This reconstruction is not arbitrary: As Scholz and Vosgerau (2025) argue, when interaction with the environment is disciplined by scientific methods – not evolutionary needs – agents can discover *methodologically objective* natural kinds. In short, a suitable structural-resemblance constraint can be developed within Gärdenfors's weakly realist framework.

Notably, T&G cite Shepard (1994) for suggesting "the general constraint that consequential regions for basic natural kinds should correspond to connected subsets of psychological space" (631), but they do not realize the significance of this issue for interpretability. Again, this relates to a point made earlier: Problems arise if geometry, or geometric constraints such as connectedness and convexity are only seen as elegant, mathematical tools in service of formal rationality. If, by contrast, they are paired with intrinsic representation *and* grounded in cognitive economy, then opaqueness can be avoided.

To summarize, Poth's similarity-based solution comprises three interconnected commitments:

(i) Commit to the presupposed similarity measure.
(ii) Commit to intrinsically meaningful psychological entities.
(iii) Ground those entities in cognitive economy.

[10] I thank Nina Poth for insightful comments on the distinction between rationality and optimality and for many additional suggestions that improved my understanding of the issues discussed here.

[11] Other authors use the non-directional notion of isomorphism, (cf. Shepard., 1981 and Edelman, 1998).

Item (i) addresses O1; item (ii) addresses O2; and item (iii) adds an additional constraint to "render learning from experience as action-guiding" in a semantically transparent way. Poth addresses (i) and (ii) via CS; a CS-based route to (iii) is feasible if Douven and Gärdenfors's optimality criteria secure the requisite connection to the world – which, I argue, they do. None of this is to say that CS are the only solution for semantic deficiency. Tversky's set-theoretic account of similarity may help with (i), but suffer from (ii) due to its symbolic nature. The goal is not to discuss all options, but to investigate one particularly promising option.

3.2 Hybrid Models

The CS account, however, does not privilege probabilistic reasoning, and Gärdenfors is relatively skeptical of Bayesian cognitive science. Yet, several authors combine CS with Bayesian reasoning, showing a degree of compatibility. Brössel (2017) discusses perceptual versus conceptual content; Scholz (2026) addresses Predictive Processing; and Strößner (2022b) develops a cognitive-naturalness criterion that harnesses geometric and probabilistic information. I focus here on three of the most developed proposals: Douven et al. (2025), Sznajder (2017, 2021), and Poth (2019). All these authors offer compelling reasons for complementing Bayesian reasoning with CS and vice versa, which I endorse because pure Bayes is semantically deficient, and pure CS does not easily account for probabilistic reasoning and related phenomena.

The hybrid accounts to be introduced define probability distributions *over* CS – a move which is inspired by Carnap's (1971, 1980) notion of attribute spaces, which are a forerunner of CS. Carnap's γ-rule "states that the prior probability $P(Ba)$ for observing an object falling under any given predicate B, should equal the normalized size ('width') of the region corresponding to that predicate in the attribute space" (Sznajder 2021, p. 469). The rule is an instance of the *principle of indifference*, according to which all options receive the same prior, if no other information is available. Decock et al. (2016) use CS to make the rule precise. Formally, to calculate the size of a region $C \subseteq S$ in a continuous n-dimensional space S, define the indicator function:

$$\mathcal{I}_C(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } (x_1, \ldots, x_n) \in C \\ 0, & \text{otherwise} \end{cases}$$

The Lebesgue-measure $\mu(C)$ of C is then:

$$\mu(C) = \int_S \mathcal{I}_C(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

To obtain a probability rather than a raw measure, normalize by dividing by $\mu(S)$. Douven et al. (2025) refine the measure to accommodate prototypes, graded membership, and fuzzy boundaries, but those details are not essential here (see also Douven et al. 2013). More importantly, they extend the account to cover iterative belief state updates and analogy influence. Let x be a sequence of point observations $\langle o^1, \ldots, o^n \rangle$ in the space, the update is a weighted Gaussian mixture:

$$\pi_i(x) = w\pi_{i-1}(x) + (1-w)d_i(x)$$

, where $w \in [0,1]$ is a weighting parameter expressing inductive inertia (the higher the w, the more weight is retained from past belief states), and $d_i(x)$ is a Gaussian distribution centered at the new observation x_i. The distribution is modulated by a spread parameter that "leaves open the possibility that the analogy influence […] is different along the […] dimensions." (Douven et al. 2025, p. 12). Then, for any concept C representable in S, the probability that the next observation $x^{i+1} \in C$ given x is calculated by integrating the current belief distribution over C:

$$Pr\left(x^{i+1} \in C\right) = \int_C \pi_i(x)\, dx$$

This model has two layers: a rigid space with a predefined conceptual partitioning, and a continuously updated probability distribution. Although it doesn't utilize Bayes' theorem directly, it is Bayesian in the sense that it models belief update by posterior probability. Its chief aim is to capture analogy influence, the idea being that observations should not only confirm the predicates they instantiate but also their similar neighbors – which Carnap modeled by a Gaussian distribution called the η-curve.

Poth (2019) instead extends the few-shot learning account by T&G. She retains the generalization and Bayesian-updating functions but formulates a CS version of the size principle by defining hypotheses as mappings from CS regions C_i to linguistic labels L and construing confirmation as overlap between evidence x_i and C_i. With region sizes measured by the Lebesgue measure, the likelihood is

$$Pr\left(x_i | H_{\langle C_i,\ L\rangle}\right) = \frac{\mu\ (x_i \cap C_i)}{\mu\ (C_i)}$$

.

Sznajder's (2017) architecture also resembles T&G in using a distribution over a hypothesis space Θ – only the hypotheses are not mappings from regions to labels, but unimodal beta distributions over a one-dimensional CS (the interval $[0,\ 1]$). In other words, they are hypotheses about what distribution is responsible for the observations. The distributions are shaped by a beta parameter, and indexed by a continuous parameter $\theta \in \Theta$ giving the mode's location. Given observations $x^{1:n}$, the posterior is:

$$Pr\left(h_\theta | x^{1:n}\right) = \frac{Pr\left(x^{1:n} | h_\theta\right) Pr\left(h_\theta\right)}{\int_\Theta Pr\left(x^{1:n} | h_\theta\right) Pr\left(h_\theta\right) d\theta}$$

If the observations are points, their likelihood $Pr\left(x^{1:n} | h_\theta\right)$ is obtained by evaluating the beta density at each point. If observations are intervals (uncertain regions), the likelihood integrates the beta density over those intervals. Note that Sznajder aims

for a fully general account of sequential prediction on CS, while analogical reasoning and perceptual generalization fall out if the hypotheses are Gaussian distributions.[12]

3.3 Discussion

Poth's model is designed to meet her desiderata: it answers O1 and O2 by committing to the similarity measure and intrinsically meaningful psychological entities (geometrically structured concepts), thereby avoiding both the *just-so* trap and the symbol-grounding problem. The same holds for the other models discussed here. Even so, Douven et al. (2025) write that they "have only been concerned with the computational level in Marr's (1982) sense (what gets computed) and not with the algorithmic level (how it gets computed), let alone with questions about hardware implementation. Algorithmic-level questions are hard for most Bayesian, or more generally probabilistic, approaches to cognition" (p. 15). Sznajder (2017), following Carnap, explicitly develops a *normative* inductive logic – a proposal about how scientists should reason. Because she does not use CS to explain how agents in fact reason, Poth's criticism does not directly apply. By contrast, Douven et al. aim to explain real agents and provide empirical corroboration. They do not invoke the computational level in a way that renders their account semantically opaque: they acknowledge that real agents likely only approximate the computations they specify, treating their algorithm as a goal multiply realizable by various approximations. At the same time, they commit to specific, intrinsically meaningful representations and to processes that operate on them. This suggests that distinctions among explanatory levels are not always sharp. The practical moral is that modelers should make explicit how computational claims are constrained by assumptions from lower levels – precisely what T&G do not do, leaving them vulnerable to O1.

The border between normative and descriptive models of cognition is likewise not sharp (see Strößner 2023). That need not detain us. Even if Sznajder's account is purely normative, it retains considerable explanatory potential for descriptive research because of its semantic transparency. When I criticize her account or suggest improvements, note that my aims differ from hers.

All the models considered here operate on *static* spatial structures: the probability distribution evolves, while the underlying space remains fixed. Real agents, however, continually acquire conceptual knowledge, which should reshape their representational frameworks (see also Strößner, 2022a). Gärdenfors discusses concept acquisition (2019) and scientific theory change (Gärdenfors & Zenker 2010, 2013), but a clear account of the formal mechanisms driving CS dynamics is still lacking. More importantly for present purposes, it remains largely unclear whether and how probability functions in hybrid models should *shape* the CS over which they range. The models above do not address this – except for Sznajder's (Sect. 4.7) discussion of

[12] In her 2021 work, Sznajder generalizes her approach using Bayesian non-parametric techniques. She proposes using a Dirichlet process prior, an infinite-dimensional prior over distributions. A Dirichlet process can be thought of as a distribution on the space of all probability distributions on a continuum. This allows very flexible hypothesis spaces, and affects some of the math, but the outlines of her model remain intact, so I do not discuss this extension here.

predicate-hypothesis mismatches.[13] Her model allows cases in which an observation falls under one predicate yet confirms another to a greater degree – e.g., when it supports a distribution hypothesis whose mode lies over the second predicate. To avoid such violations of Carnap's self-similarity principle (roughly: each predicate most resembles itself, and the analogy function should capture this), she proposes placing category boundaries where distribution hypotheses intersect. This tells us how the hypothesis distributions affect partitioning, but not how the degrees of belief reshape the underlying spatial structure. She tentatively suggests that the conceptual partition should track the empirical distribution in hypothesis space – subdividing dense areas into more specific concepts – which echoes the idea that natural concepts carve nature at its joints, placing boundaries where data density drops (Strößner 2022b). She does not, however, supply a mechanism. Moreover, our primary interest here is not partitioning but the dimensions of the underlying space.

As Frigg and Hartmann (2025) emphasize, useful idealization is a virtue of scientific models. It would therefore be unfair to deem Douven et al. or Poth cognitively inadequate for not modeling CS dynamics (and Sznajder does not aim for cognitive adequacy at all). Nevertheless, an account of the dynamics of representational geometry would enhance the cognitive adequacy of CS models in general. More specifically, an account of how probabilistic information extracted from observations reshapes the dimensions would substantially advance hybrid models. The remainder of the paper outlines such an account and discusses whether it is semantically deficient.

4 Application

To recap: semantic deficiency is a significant problem, CS plus Bayesian reasoning hybrids are a solution, but existing models exhibit static spatial structure. The model to be sketched here addresses this research gap by using the degrees-of-belief function to compute higher-order dimensions in a hierarchical model, thereby capturing *abstraction* in CS.[14] It serves as both an application of the insights from the discussion so far, and a second case study – T&G being the first – for the notion of semantic deficiency. The proposed mechanism relies on principal component analysis (PCA), a widely discussed statistical tool for capturing covariation. What is new is to use the belief function to generate a sample of synthetic observations that serve as input to PCA. A potential disadvantage is that PCA-based abstraction may introduce semantically opaque dimensions.

[13] Note that this also holds for Brössel's, (2017) model, which I do not discuss here because it is a proof of principle rather than a general model of rational belief update. Brössel represents color experiences as points in a phenomenal similarity space and color concepts as regions in that space, and then uses probability-theoretic machinery to show how a perceptual belief can be *justified* given a perceptual experience. This creates a probabilistic bridge between non-conceptual perceptual contents and perceptual beliefs, but it does not capture how the probability function, or updates to it, might reshape the structure of the conceptual space itself – e.g., its dimensions or metric.

[14] Abstraction is understood here as an omission of detail in the service of cognitive economy. The proposed mechanism omits detail by compressing high-dimensional data into lower-dimensional spaces.

Before detailing the mechanism in the next subsection, note why abstraction matters, why PCA is attractive, and why a hierarchical model is appropriate. First, Gärdenfors (2000) positions CS as a bridge between connectionist and symbolic theories; abstraction provides the pillars of this bridge. Cognitive systems are assumed to distill the *blooming*, *buzzing confusion* (a phrase coined by William James) of high-dimensional sensory input into a manageable set of meaningful dimensions, yielding CS. According to Gärdenfors (2019), two steps are involved: (1) Perceptual domains are constructed by extracting invariants, e.g., shape is invariant under location transformations. (2) "Once the primary domains are in place, the brain is efficient in finding covariances of different features" (p. 460), e.g., the sweetness and the color of strawberries systematically vary together. PCA is a natural candidate for dimensionality reduction that detects and leverages covariance. Finally, many concepts are hierarchically organized (e.g., *Amanita phalloides* $\subset$ *Amanita*), and hierarchical spaces connect CS with Predictive Processing (PP), a Bayesian-inspired framework that posits hierarchically organized probabilistic models minimizing prediction error (cf. Scholz 2026). The current proposal addresses bottom-up abstraction; top-down prediction is a natural extension for future work.

4.1 PCA-Based Abstraction ($S \rightarrow S\prime$)

PCA reduces the number of variables (dimensions) in data while preserving as much meaningful variation as possible (Jolliffe 2002). It is effective when features are correlated. Principal components (PCs) are new orthogonal directions in the data which "explain" variance by capturing these correlations. This way of speaking is justified insofar as PCs support prediction of future data – recall Schurz (2014). The first principal component captures most of the variance, the second one is uncorrelated with (orthogonal to) the first and captures most of the variance that is left, and so on. Because PCA is a dimensionality reduction technique, it has attracted attention in the CS literature. Gärdenfors and Osta-Vélez (2026) use it to define *conceptual coherence*. For highly coherent natural kind concepts such as BIRD, "multiple correlated dimensions – like *size*, *beak shape*, *habitat*, and *diet* – […] systematically covary" (p. 210), so few PCs capture most of the variation. They suggest PCA can cash out the natural-kind/artifact contrast (cf. Schwartz 1978) because artifacts are less coherent, proposing that the first PC(s) may characterize a concept's "essence" in a psychological sense and reducing the latter to covariation (Gärdenfors & Osta-Vélez, submitted, p. 10). Accounting for psychological essentialism is an independent motivation for PCA beyond the opaqueness question.[15]

Two issues matter when adapting PCA to hybrid models. First, standard practice "assumes all dimensions are equally important, since it standardizes each variable's scale before computing correlations" (Gärdenfors & Osta-Vélez 2026, Sect. 9.3.3). That is implausible for conceptual content, where some dimensions are more *salient*

[15] When the authors speak of PCs as "essences" or "causes", we must not lose sight of the fact that they are condensed portrayals of data. Thus, we are dealing here with a reductive, or deflationary view of these concepts. Discussing the metaphysical and broader philosophical ramifications of this view is beyond the scope of this paper.

than others (e.g., gill presence may matter more than cap diameter for mushroom categorization). The authors "propose weighting covariations according to salience rather than relying on correlations derived from standardized variable ranges [to ensure] that the principal components better reflect the structure of the concept rather than merely reflecting statistical variance" (ibid.). We follow the authors' proposal while leaving the weighting implicit. Second, PCA takes vectors, not probability distributions, as inputs. We therefore draw a number of *synthetic samples* from the belief distribution and feed those to PCA. While this adds a step, it is arguably cognitively economical: agents encounter many stimuli over time; instead of storing each, they can use the current belief distribution to generate a plausible sample.

Let the base conceptual space be $S \subseteq \mathbb{R}^n$ with coordinates $x = (x_1, \ldots, x_n)$. After i observations, the agent maintains a belief distribution π_i on S (e.g., the weighted Gaussian mixture of Douven et al. 2025). Draw N synthetic instances:[16]

$$x^1, \ldots, x^N \sim \pi_i$$

Applying PCA (with salience weighting instead of standardization) to $\{x^{(j)}\}$ yields orthonormal directions $v_1, \ldots, v_m \in \mathbb{R}^n$ for $m < n$ that capture the dominant covariance in the sample. These directions serve as the quality dimensions of $S\prime$. PCA simultaneously "allow[s] the representation of data elements of A in a representation space B of lower dimensionality, while maintaining similarity relations in A" (Kaipainen & Hautamäki 2015, p. 250). In other words, points can be projected from S to S', which enables their categorization, etc. in the abstract space.[17]

Take, as an example, the three-dimensional strawberry space comprising sweetness, color and size. Fictive samples from π_i will typically form an elongated cloud from unripe to ripe (viz., green-and-sweet strawberries are very unlikely, and so forth). Performing PCA on the samples yields a first, abstract dimension which explains most of the variation, and which can be interpreted as ripeness. Due to the projection, the observations are no longer just elements of SWEET, etc., but can be categorized on more abstract level as RIPE, given a fitting conceptual boundary on the ripeness dimension.

Both the sample size N and the number of PCs m are left open and are, again, governed by cognitive economy. Larger sample size leads to a better fit between

[16] Drawing random samples from a probability distribution in order to approximate quantities of interest is an instance of *Monte Carlo sampling* – an established method in the Bayesian literature. See (Aslett, 2022).

[17] Like the formal details of the PCA step, we keep the projection implicit in the model. Intuitively, to locate any $x$in $S\prime$, measure how far $x$lies along each new quality dimension. Kaipainen and Hautamäki (2015) develop a *perspectival* projection step which could be plugged in. Their account of abstraction is comparable to the one presented here, but they approach the issue from a very different perspective and do not discuss Bayesianism.

$x$in

$S\prime$, measure how far

$x$lies along each new quality dimension. Kaipainen and Hautamäki (2015) develop a *perspectival* projection step which could be plugged in. Their account of abstraction is comparable to the one presented here, but they approach the issue from a very different perspective and do not discuss Bayesianism.

the distribution and the abstract dimensions, but is cognitively costly. One plausible factor to determine the sample size could be the number of modes of the distribution, since more modes means more complexity, which requires more detail. The number of PCs should depend on how much of the variation is explained by each PC. Intuitively, one PC suffices in the strawberry example, but the value of an economic threshold will depend on the data, and perhaps on the context and goal of the abstraction. Since abstraction is supposed to be an omission of detail, however, we set $m < n$.

Four open design choices shall be highlighted: (i) *Abstract-level belief distribution*. One could induce a distribution on $S\prime$ by projecting π_i, analogous to coordinate projection; I defer this. (ii) *Recursion*. The two-step procedure can be iterated $S \rightarrow S\prime \rightarrow S\prime\prime$ to build deeper hierarchies; I do not pursue this here. (iii) Salience pre-scaling. With salience weights $s_1, \ldots, s_n$, dimensions can be pre-scaled before PCA. In models with top-down influence, salience could be linked to each lower-level dimension's contribution to the PCs. (iv) *PCA recalculation*. As experience accumulates, π_i changes. If abstract dimensions cease to capture the distribution, recalculating PCA may be warranted. Determining triggers for recalculation is an important question for future work.

The proposal does not collapse probabilistic into geometric information. Rather, it uses probability as a heuristic to generate new dimensions. Similarity and probability remain distinct, but the mechanism hypothesizes how they might interact in the human mind. It should be pointed out, however, that one may doubt whether the proposed mechanism is psychologically realistic. While the idea is that it is economic not to store every piece of evidence, but generate a realistic sample, probability distributions themselves are standardly regarded as computationally intractable for biological systems, which is why sampling methods are often used to approximate them. It thus remains an open question what cognitive process corresponds to sampling from one's own degree of belief. I set these problems aside here to focus on the question whether the proposed model satisfies EC.

4.1.1 Reintroducing Semantic Deficiency?

PCA detects directions in data; it does not label them. The same holds for other dimension-inducing techniques (e.g., multidimensional scaling; see Gärdenfors 2000; Sect. 1.7). Often, reasonable interpretations exist (e.g., ripeness), but they remain interpretations. One might worry this invites *relativism*, given the subjective nature of interpretation (Gärdenfors 1990, Sect. 6). Furthermore, it has been noted that CS works best for perceptual concepts (e.g., color) and struggles with abstract ones (e.g., intuitively, it is hard to tell what the dimensions of a space with a JUSTICE-region could be); technical work addresses metrics and geometric naturalness for abstract concepts (Strößner 2022b). What is missing is an analysis of how abstraction bears on semantic deficiency. Consider the following examples:

(i) *Too fine-grained (sub-conceptual):* olfaction. Let the base space be high-dimensional olfactory receptor responses. PCA on fictive samples yields mixtures of many receptor channels. These axes robustly capture variance, yet they fail to

align with any recognizable odor quality ("minty," "smoky," etc.). They are directions in a mixed sensor space, not nameable properties – hence practically uninterpretable.

(ii) *Too coarse-grained (hyper-abstract):* human flourishing. Let the base space combine indicators often used to talk about "flourishing" (income, health, social connection, autonomy, environmental quality, etc.). PCA's first axis collapses various indicators into a single "development bundle" (e.g., wealth, education and urbanization). It detects covariations but does not isolate nameable flourishing facets (e.g., hedonic, eudaimonic, etc.). The abstraction is above the level at which our predicates apply.

(iii) *Spurious features*: fish identification. Let the base space consist of features of fish specimens in trophy photos (shape ratios, fin geometry, texture, color). Fictive sampling and PCA yield a leading axis that captures covariances related to anglers (skin tone, hand contours, etc.), not to the fishes themselves.

All of these examples invite suspicion that PCA leads to semantic deficiency. But O1 and O2 are not the issue this time because this is a hybrid model with intrinsically meaningful dimensions at the base level, and so forth. The abstraction step is what fails, but in different ways. Examples (i) and (ii) are driven by a granularity mismatch: In (i), axes sit below the level where we have linguistic labels, and in (ii), axes sit above the level where we have linguistic labels. Accordingly, semantic features are present, but there is no way for linguistic labels to attach.[18] Recall that Facchini and Termine (2022) speak of semantic opacity if model interpretation is impaired – either because there is no well-defined semantics, or the users lack the skills to understand it, or it is inadequate for the context. O1 arises from the first condition, while (i) and (ii) primarily relate to the second, and possibly the third condition. Example (iii) differs: the axis is interpretable, but the interpretation doesn't match the target ontology. The example is inspired by the *spurious feature* literature on connectionist models. Spurious features are features in the data that models detect to meet classification benchmarks, for example, but which are in some sense not the right features. Neuhaus et al. (2023) "develop a pipeline for the detection of harmful spurious features with little human supervision based on [their] class-wise neural PCA (NPCA) components of an adversarially robust classifier together with their Neural PCA Feature Visualization (NPFV)" (p. 2). That is, they use PCA to extract the components which are responsible for the classification behavior, and visualize them to judge whether they are the right features.

Such post-hoc methods, however, are needed for connectionist models, but not for the examples. Indeed, case (iii) is not an instance of opaqueness; it shows how PCA can help *avoid* problematic opaqueness. In the example, the fact that angler features constitute the leading PC is perfectly transparent. This is plausible because PCA provides a straightforward way to tell how much each dimension contributes to each component: We can think of PCs as weighted combinations of the original dimensions. Calculating the weight of each dimension – the so-called "loadings"

[18] I assume here that concepts can exist without linguistic label (viz., non-human animals can have concepts), but that linguistic labels presuppose concepts.

– is a regular step of any PCA (cf. Abdi & Williams 2010, p. 438). As the original dimensions are interpretable from the start, spurious features are easy to spot via the loadings. The same mechanism can also be used to get a sense of hyper-abstract components, even when they cannot be named. It does not solve (i): when base-space features are already too low-level for linguistic labels, PCs inherit that limitation until a certain level of abstraction is reached. Until then, a form of opacity remains which is close to the semantic opacity of connectionist systems.

This residual opacity in ii) is harmless because explanatory adequacy in cognitive science requires showing how cognitive content produces behavior (EC). The models from ii) and iii) do that via PCA-loadings, even if linguistic labels cannot attach to the contents – which is unavoidable for some levels of abstraction – or if they transparently pick up on the wrong contents. Thus, PCA-based abstraction may introduce dimensions which are opaque in the sense that no linguistic labels attach. Semantic deficiency only arises for case (i), but this is due to connectionism, and not to the PCA-abstraction.

5 Conclusion

This paper has examined connections among Bayesianism, Conceptual Spaces (CS), AI opacity, and explanation in cognitive science. It has highlighted a facet of opacity that remains poorly understood: semantic opacity in the context of modeling cognitive capacities. This is a theoretically interesting context which deserves more attention in the XAI literature. The discussion has also revealed that many Bayesian models of cognition leave psychological assumptions implicit – paradigmatically in T&G, assumptions about how similarity structures the hypothesis space – thereby risking just-so stories (O1). Because they also share features with GOFAI systems, they face the symbol-grounding problem (O2). While the latter has nothing to do with opacity, it also frustrates the explanatory constraint EC, which is thus revealed to be the thread connecting O1, O2, and semantically opaque (in the sense of Facchini & Termine) models of cognition. To avoid confusion, it was suggested to drop Poth's original terminology and speak of "semantically deficient models of cognition" (SEDMOC) instead.

CS can address O1 by providing an explicit similarity structure for the hypothesis space, and O2 by supplying intrinsically meaningful inputs to Bayesian machinery. We have argued that it can also provide a rational connection with the world by grounding sparse concepts in optimality. The discussed hybrid models meet EC by combining CS with probabilistic reasoning. Like most scientific models, they are idealizations; nevertheless, articulating how the underlying space changes over time – and how the probability distribution should influence such changes – would enhance their cognitive adequacy. The final section outlined a two-step procedure: draw plausible samples from the belief function and apply PCA to generate abstract dimensions in a hierarchical model. Although PCA can yield dimensions that are hard to name or that sometimes pick out the wrong features, this does not necessarily lead to semantic deficiency, as PCA loadings make transparent which cognitive contents contribute to each component, so the explanatory constraint derived earlier is preserved. Of

course, how much weight one gives to the dimensions having natural language interpretations may again be a matter of pragmatic research interests and context – semantic deficiency is not meant as an absolute, but as a context-sensitive term.

Further work should clarify how top-down influence operates in hierarchical CS and how probability distributions behave across multiple layers. The notion of semantic opacity itself also merits development. Facchini and Termine's (2022) typology is a useful orientation in a messy debate and places semantic opacity squarely on the agenda. Our analysis of how CS can avoid O1 and O2 contributes to that development, while indicating concrete paths toward more cognitively realistic hybrid models.

Acknowledgements I thank my supervisors and colleagues at the HHU for their continued input and support: Benedikt Kenyah-Damptey, Chiara Lisciandra, Julia Mirkin, Nastasia Müller, Gerhard Schurz, Maria Sojka – and especially Gottfried Vosgerau. I would also like to thank my friends and colleagues at Lund University for enabling a wonderful and productive research visit, the result of which is this paper: Andrey Anikin, Vidar Bratt, Niklas Dahl, Jiwon Kim, Pierre Klintefors, Sarah Kögelsperger, Elsa Magnell, Shervin Mirzaeighazi, Manuel Quezada, Simon Rosengren, Sophie Serra, Andreas Stephens, Trond Arild Tjøstheim, Balder Ask Zaar. Special thanks to Peter Gärdenfors and Nina Poth for their most helpful comments and suggestions.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding was provided by my institution, the Heinrich-Heine-University Düsseldorf. No third-party funding was involved.

Data Availability Not applicable.

Declarations

Competing Interests I do not have any competing interests to declare.

References

Abdi, Hervé, and Lynne J.. Williams. 2010. Principal component analysis. *WIREs Computational Statistics* 2 (4): 433–59. https://doi.org/10.1002/wics.101.

Aslett, Louis J. M. 2022. Sampling from complex probability distributions: A Monte Carlo Primer for engineers. In *Uncertainty in Engineering: Introduction to Methods and Applications*, edited by Louis J. M. Aslett, Frank P. A. Coolen, and Jasper De Bock. Springer International Publishing. https://doi.org/10.1007/978-3-030-83640-5_2.

Bermúdez, José Luis. 2014. *Cognitive Science: An Introduction to the Science of the Mind*. 2. ed. Cambridge Univ. Press.

Bowers, Jeffrey S.., and Colin J.. Davis. 2012. Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin* 138 (3): 389–414. https://doi.org/10.1037/a0026450.

Brössel, Peter. 2017. Rational relations between perception and belief: The case of color. *Review of Philosophy and Psychology* 8 (4): 721–41. https://doi.org/10.1007/s13164-017-0359-y.
Brown, Curtis. Narrow mental content. In *The Stanford Encyclopedia of Philosophy*, Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2022. https://doi.org/Meta.
Buckner, Cameron, and James Garson. 2025. Connectionism. In *The Stanford Encyclopedia of Philosophy*, Spring edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2025/entries/connectionism/.
Carnap, Rudolf. 1971. 'A basic system of inductive logic, Part I'. In *Studies in Inductive Logic and Probability*, edited by Richard C. Jeffrey. University of California Press.
Carnap, Rudolf . 1980. 'A basic system of inductive logic, Part II'. In *Studies in Inductive Logic and Probability*, edited by Richard C. Jeffrey. University of California Press.
Chater, Nick, and Mike Oaksford. 2000. 'The rational analysis of mind and behavior.' *Synthese* 122 (1–2): 93–131. https://doi.org/10.1023/A:1005272027245.
Crupi, Vincenzo, and Fabrizio Calzavarini. 2023. Critique of pure Bayesian cognitive science: A view from the philosophy of science. *European Journal for Philosophy of Science* 13 (3): 28. https://doi.org/10.1007/s13194-023-00533-w.
Decock, Lieven, Igor Douven, and Marta Sznajder. 2016. A geometric principle of indifference. *Journal of Applied Logic* 19 (2): 54–70. https://doi.org/10.1016/j.jal.2016.05.002.
Douven, Igor, Lieven Decock, Richard Dietz, and Paul Égré. 2013. Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic* 42 (1): 137–60. https://doi.org/10.1007/s10992-011-9216-0.
Douven, Igor, and Peter Gärdenfors. 2019. 'What are natural concepts? A design perspective.' *Mind & Language* 3: 313–34. https://doi.org/10.1111/mila.12240.
Douven, Igor, Steven Verheyen, Shira Elqayam, Peter Gärdenfors, and Matías Osta-Vélez. 2025. Analogical reasoning: A Carnapian approach. *Synthese* 205 (5): 1–34. https://doi.org/10.1007/s11229-025-04991-y.
Edelman, Shimon. 1998. Representation is representation of similarities. *Behavioral and Brain Sciences* 21 (4): 449–67. https://doi.org/10.1017/s0140525x98001253.
Egan, Frances. 2014. 'How to think about mental content.' *Philosophical Studies* 170 (1): 115–35. https://doi.org/10.1007/s11098-013-0172-0.
Facchini, Alessandro, and Alberto Termine. 2022. 'Towards a taxonomy for the opacity of AI systems'. In *Philosophy and Theory of Artificial Intelligence 2021*, edited by Vincent C. Müller. Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_7.
Fodor, J.A. 1975. *The Language of Thought*. Crowell.
Frigg, Roman, and James Nguyen. 2021. 'Scientific Representation'. In *The Stanford Encyclopedia of Philosophy*, Winter edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/scientific-representation/.
Frigg, Roman, and Stephan Hartmann. 2025. 'Models in Science'. In *The Stanford Encyclopedia of Philosophy*, Summer edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2025/entries/models-science/.
Gärdenfors, Peter. 1990. 'Induction, conceptual spaces and AI'. *Philosophy of Science* 57(1): 78–95.
Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. MIT Press.
Gärdenfors, Peter. 2014. *The Geometry of meaning: Semantics Based on Conceptual Spaces*. MIT Press.
Gärdenfors, Peter, and Frank Zenker. 2013. Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. *Synthese* 190 (6): 1039–58. https://doi.org/10.1007/s11229-011-0060-0.
Gärdenfors, Peter. 2019. 'From sensations to concepts: A proposal for two learning processes'. *Review of Philosophy and Psychology* 10 (3): 441–464. https://doi.org/10.1007/s13164-017-0379-7.
Gärdenfors, Peter, and Matías Osta-Vélez. 2026. MIT Press.
Gärdenfors, Peter and Osta-Vélez, Matías (submitted). *'The essence of a concept is its principal components'.*
Gärdenfors, Peter and Zenker, Frank . 2010. *Using Conceptual Spaces to Model the Dynamics of Empirical Theories*. https://doi.org/10.1007/978-90-481-9609-8_6.
Goodman, Nelson. 1983. *Fact, Fiction, and Forecast*. 4th Edition. Harvard University Press.
Hahn, Ulrike. 2014. The Bayesian boom: Good thing or bad? *Frontiers in Psychology* 5: 765. https://doi.org/10.3389/fpsyg.2014.00765.
Harnad, Stevan. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42 (1): 335–46. https://doi.org/10.1016/0167-2789(90)90087-6.

Hutto, Daniel, and Ian Ravenscroft. 2021. 'Folk psychology as a theory'. In *The Stanford Encyclopedia of Philosophy*, Fall edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/fall2021/entries/folkpsych-theory/.

Jolliffe, I. T. ed. 2002. Principal component analysis for special types of data. In *Principal Component Analysis*, Springer. https://doi.org/10.1007/0-387-22440-8_13.

Jones, Matt, and Bradley C.. Love. 2011. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences* 34 (4): 169–88. https://doi.org/10.1017/s0140525x10003134.

Kaipainen, Mauri, Antti, and Hautamäki. 2015. 'A perspectivist approach to conceptual spaces'. In *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation*, edited by Frank Zenker and Peter Gärdenfors. Synthese Library. Springer International Publishing. https://doi.org/10.1007/978-3-319-15021-5_13.

Lipton, Zachary C. 'The mythos of model interpretability'. *Communications of the ACM* 61, no. 10 (2018): 36–43. https://doi.org/Meta

Margolis, Eric, and Stephen Laurence. 2021. 'Concepts'. In *The Stanford Encyclopedia of Philosophy*, Spring edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/spr2021/entries/concepts/.

Marr, D. 1982. *Vision: A Computational Investigation in the Human Representation of Visual Information*. Freeman.

Morrison, Margare, and Mary S. Morgan. 1999. 'Models as mediating instruments'. In *Models as Mediators: Perspectives on Natural and Social Science*, edited by Margaret Morrison and Mary S. Morgan. Ideas in Context. Cambridge University Press. https://doi.org/10.1017/CBO9780511660108.003.

Neuhaus, Yannic, Maximilian Augustin, Valentyn Boreiko, and Matthias Hein. 2023. 'Spurious features everywhere -- large-scale detection of harmful spurious features in imageNet'. arXiv:2212.04871. Preprint, arXiv. https://doi.org/10.48550/arXiv.2212.04871.

Newen, Albert, and Gottfried Vosgerau. 2020. 'Situated mental representations: Why we need mental representations and how we should understand them'. In *What Are Mental Representations?* by Albert Newen and Gottfried Vosgerau. Oxford University Press. https://doi.org/10.1093/oso/9780190686673.003.0007.

Pernu, T. (unpublished manuscript). 'Is There Causation in Physics?

Pitt, David. 2020. 'Mental representation'. In *The Stanford Encyclopedia of Philosophy*, Spring edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/mental-representation/.

Poth, Nina. 2019. Conceptual spaces, generalisation probabilities and perceptual categorisation. In *Conceptual Spaces: Elaborations and Applications*, ed. Peter Gärdenfors, Antti Hautamäki, Frank Zenker, and Mauri Kaipainen. Springer Verlag.

Poth, Nina. 2023. 'Probabilistic Learning and Psychological Similarity'. *Entropy* 25 (10): 1407. https://doi.org/10.3390/e25101407.

Scholz, Sebastian. 2026. 'Conceptual Spaces as Predictive Models'. *Synthese* 207 (6): 237. https://doi.org/Meta

Scholz, Sebastian, and Gottfried Vosgerau. 2025. Conceptual spaces: Naturalness or cognitive sparseness? *Synthese* 205 (3): 129. https://doi.org/10.1007/s11229-025-04941-8.

Schurz, Gerhard. 2014. *Philosophy of Science: A Unified Approach*. Routledge.

Schwartz, Stephen P.. 1978. Putnam on artifacts. *Philosophical Review* 87 (4): 566–74. https://doi.org/10.2307/2184460.

Shepard, R.N. 1994. Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review* 1 (1): 2–28. https://doi.org/10.3758/BF03200759.

Shepard, Roger . 1981. 'Psychological complementarity'. In *Perceptual Organization*, edited by Michael Kubovy and James R. Pomerantz. Routledge. 279–341.

Shepard, Roger . 1987. 'Toward a universal law of generalization for psychological science'. *Science (New York, N.Y.)* 237(4820): 1317–23. https://doi.org/10.1126/science.3629243

Sober, Elliott. 1999. The multiple realizability argument against reductionism. *Philosophy of Science* 66 (4): 542–64. https://doi.org/10.1086/392754.

Strößner, Corina. 2022a. Conceptual learning and local incommensurability: A dynamic logic approach. *Axiomathes* 32 (6): 1025–45. https://doi.org/10.1007/s10516-021-09563-6.

Strößner, Corina. 2022b. 'Criteria of naturalness in conceptual spaces'. *Synthese* 200(78): 36.

Strößner, Corina. 2023. 'Hybrid frameworks of reasoning: Normative-descriptive interplay'. *Proceedings of the Annual Meeting of the Cognitive Science Society* 45(45). https://escholarship.org/uc/item/58j925zr.

Sullivan, Emily. 2022. Understanding from machine learning models. *British Journal for the Philosophy of Science* 73 (1): 109–33. https://doi.org/10.1093/bjps/axz035.

Sznajder, Marta. 2017. *Inductive Logic on Conceptual Spaces*. Rijksuniversiteit Groningen.

Sznajder, Marta. 2021. 'Inductive reasoning with multi-dimensional concepts'. *British Journal for the Philosophy of Science* 72(2): 465–484. https://doi.org/10.1093/bjps/axz012.

Tenenbaum, Joshua B.., and Thomas L.. Griffiths. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences* 24 (4): 629–40. https://doi.org/10.1017/s0140525x01000061.

Williams, Daniel. 2018. Predictive processing and the representation wars. *Minds and Machines* 28 (1): 141–72. https://doi.org/10.1007/s11023-017-9441-6.

Woodward, James, and Lauren Ross. 'Scientific explanation'. In *The Stanford Encyclopedia of Philosophy*, Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021. https://doi.org/Meta.