



A general formulation of conceptual spaces as a meso level representation

Janet Aisbett*, Greg Gibbon

School of Information Technology, The University of Newcastle, Callaghan, Australia

Received 21 September 2000; received in revised form 23 April 2001

Abstract

Representing cognitive processes remains one of the great research challenges. Many important application areas, such as clinical diagnosis, operate in an environment of relative magnitudes, counts, shapes, colours, etc. which are not well captured by current representational approaches. This paper presents conceptual spaces as a meso level representation for cognitive systems, between the high level symbolic representations and the subconceptual connectionist representations which have dominated AI. Conceptual spaces emphasize orders and measures and therefore naturally represent counts, magnitudes, and volumes. Taking Gärdenfors' decade-long investigation of conceptual spaces [Gärdenfors, *Conceptual Spaces: The Geometry of Thought*, MIT Press, 2000] as start point, the paper presents a formal foundation for conceptual spaces, shows how they are theoretically and practically linked to higher and lower representational levels, and develops dynamics which allow the orbits of states in the space to solve appropriate meso level reasoning tasks. Interpretations of conceptual spaces are given to illustrate the formal definitions and show the flexibility of the representation. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Concept representation; Cognitive processing; Feature spaces; Dynamical systems; Knowledge representation; Conceptual spaces; Representational levels; Categorisation; Prototypes; Conceptual distance

1. Introduction

Representing cognitive processes remains one of the great research challenges. Consider the important area of clinical diagnosis, say, differentiating between iron deficiency and anaemia due to leukaemia [35]. Symptoms are often mild, and include common conditions such as increasing fatigue, lightheadedness, or palpitations. Clinical signs are

* Corresponding author.

E-mail addresses: mgjea@cc.newcastle.edu.au (J. Aisbett), mgggg@cc.newcastle.edu.au (G. Gibbon).

also subtle—pallor, spoon-shaped finger nails, tongue texture, heartbeat irregularities and so on. Laboratory tests look firstly at red blood cell counts and average volumes, then at cell colour and shape. How well can this environment of relative magnitudes, shapes, colours and counts be captured with current representational approaches?

Symbolic representations have dominated in application areas such as these, from the early expert systems through to the current activity in medical ontologies and terminological systems like the UMLS semantic network [32]. Automated neural networks are applied in medical equipment monitoring, image recognition, and diagnosis systems where, as with symbolic approaches, input is usually mid level feature vectors. The anaemia differentiation problem would be amenable to either type of approach only after extensive cognitive effort had been put into extracting and abstracting information.

This paper builds on Gärdenfors' [15] development of conceptual spaces as a mid level representation for cognitive systems, between connectionist and symbolic systems. Conceptual spaces, as defined by Gärdenfors, emphasize orders and measures and therefore naturally represent counts and magnitudes, which are so much of the fabric of cognition. The paper's contributions are

- (a) a formal foundation for conceptual spaces,
- (b) a demonstration of how conceptual spaces are theoretically and practically linked to higher and lower representational levels, and
- (c) the development of dynamics which allow states to represent complex structures, and the orbits of states to solve appropriate “meso level” reasoning tasks, including categorisation.

1.1. Cognitive systems

Representations are sometimes defined to be substates of a cognitive system that support the system's purposeful interaction with its environment (e.g., [11,26]). Although representations may have physical realisations, they are not required to have them: that is, representation may be an abstract modelling tool used to *describe* mental activities of organisms or activities of artificial systems. Whether just a descriptive tool, or whether physically realised, representations are goal-directed abstractions of an environment which can itself be natural or constructed, or a combination.

For the system's interactions with its environment to be non-accidental, the state of the external environment must be able to affect the state of the system. That is, there must be some *perceptual input* mechanism, as when the clinician observes the anaemia signs. For the achievement of the goals of the system to be linked to the state of the environment, these goals must depend on *differentiation* amongst states and *identification* of states. So the representations must make distinctions between relevant states or sets of states of the environment.

Since the physical environment is comprised of (externally) measurable forces and constituents—heart rate, average size of red blood cells, aircraft speed, number of wings etc.—the system must be capable of approximating both magnitudes and counts in its representations. In biological and artificial systems this occurs in the accumulated activity of units in selected subsystems, in the case of magnitude, and in the count of selected active subsystems, in the case of count (e.g., [10]). A very direct example of representation

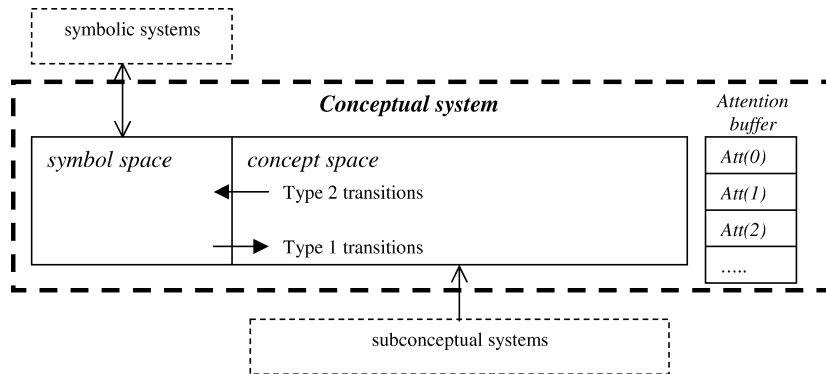


Fig. 1. Schematic of conceptual spaces as a meso level representation form, which can be viewed as having inputs from “lower level” (subconceptual) and “higher level” (symbolic) systems, analogously to meso level representations in physics. The structuring of conceptual spaces into symbol and concept spaces with an attention buffer is described in Section 4. Symbol space is described there as being a vector space, where each component of a vector represents the activity of a “symbol” or label. The dynamics of the meso level system, including the setting of the attention buffer and the type 1 and type 2 state transitions, are described in Section 5.

of magnitudes is provided in van Gelder and Port’s [42] description of Watt’s steam engine governor as a system with goal-directed interaction with the environment: here, the environmental measurable of pressure determines the state of the system.

This paper considers systems that also interact with other systems through a symbol-based subsystem, as the clinician does talking with the patient. A pressure setting and display device would be an example of such a subsystem with which Watt’s governor could communicate with humans. For the interaction to be meaningful, the system must be dependable in the way its internal states are associated with the symbols or tokens it presents externally, and in its reactions, given its internal state, to the symbols it receives from the external world. The system may also have a memory in which to store sequences (programs) of commands to direct attention to parts of its representations, so that a single change in the symbol-based subsystem can result in a sequence of operations to transform the system’s state. Fig. 1 provides a schematic of the system described in this paper.

Cognitive research has been heavily influenced by, first, symbolic and, then, connectionist representations (or more generally, by subconceptual representations as dynamical systems). As Gärdenfors points out, pure symbolic systems operate with a realist semantics that presume external representations. Mappings to the external world are needed to provide symbols with their relationship to the environment. Without such mappings, for example, magnitude and count are only formally available through infinite constructions, although they are used in quantities, probabilities and the like, through informally co-opting arithmetic. On the other hand, connectionist representations focus on processing input at the perceptual level and generally have weak links between the perceptual input and the symbolic level. Artificial neural net systems designed for high level output usually have input which must be interpreted at a high level, e.g., classification systems for medical diagnosis.

Some connectionist systems distinguish two sets of nodes, in which the nodes of one set, call it the higher level, are considered to be symbols which are associated with patterns of activity in the second set (e.g., [5,29]). They can be interpreted as implementations of the conceptual systems described below. However, they are but examples, without an equivalent theory based on geometry. There has been a theoretical gap between the two main AI representational forms, symbolic and connectionist. The fact that systems are rarely if ever *implemented* as pure connectionist or pure symbolic systems—because, for example, they incorporate digital arithmetic units or electronic sensing devices, or because they are hybrid systems—encourages the development of the meso level theory.

1.2. *Structure of cognitive representations*

A recurring feature of cognitive mechanisms used to differentiate and identify states of the environment is the assignment of *properties* to sets of substates variously called *objects* and *concepts*. Kirsch says: “*It means that the creature is able to identify the common property which two or more objects share and to entertain the possibility that other objects also possess that property. That is, to have a concept is, among other things, to have a capacity to find an invariance across a range of contexts, and to reify that invariance so that it can be combined with other appropriate invariances*” [24].

Dimensions and domains form the framework used to assign properties to concepts and objects, and to specify relations between them. Because basic concepts are not always independent of each other, interdependent concepts are usefully organised into *domains*. Domains might be derived from the perceptual mechanism primarily responsible for the information, from associations learned through feedback from the environment, or might be ascribed internally by the system. Thus colour concepts belong to one domain, concepts for sounds to a second, company ownership relations to a third, and so on. There is ample support from neurophysiology and neuropsychology for domain-specificity in the brain, whether innate or built up through experience (e.g., [33,41]).

Domains in turn are composed of *dimensions*, the primary function of which is to represent various “qualities” of situations or objects. Examples of quality dimensions are sensory-derived qualities such as temperature or the three ordinary spatial dimensions of height, width and depth; qualities of an abstract non-sensory character, such as integrity or popularity; or internally derived qualities such as level of fatigue or fear.

Differentiation and identification require judgements about the similarity of states. In the context of such judgments, Shepard and Chipman [37] say of representation that “*isomorphism should be sought—not in the first order relation between (a) an individual object and (b) its corresponding internal representation—but in the second order relation between (a) the relations among external objects, and (b) the relations among their corresponding internal representations*”. Relations amongst representation of objects and concepts will be determined by betweenness and distance relations on quality dimensions, as discussed in Section 3.

A conceptual space is in essence a multidimensional space in which the dimensions represent qualities or features of that which is being represented. A point in the space is a state in the associated conceptual system. Gärdenfors’ primary motivation for introducing conceptual spaces was to provide new tools for manipulating explicit representations of

concepts. Geometry provides representational capacities and analysis tools not naturally available in the connectionist and symbolic representational forms. Some phenomena, such as the symptoms, signs and tests of anaemia, are more easily and accurately modelled at the conceptual level. Although geometry has been exploited in specialised representations, instance based learning for example, the vector space structure often assumed is neither necessary, nor always appropriate.

Conceptual representations can be taken to be the references of symbols, and to receive perceptually-derived input. As such, the conceptual level is naturally seen to lie between the symbolic and the subconceptual levels. The development of conceptual spaces as a generic mechanism for representation and reasoning—in partnership with these and other representational forms—may be as important as meso level modelling has proved to be in atmospheric physics, economics, neurobiology and so on.

1.3. About this paper

At the outset, it is not at all obvious how to best formulate a conceptual space or to embed objects in such a space. However, for an intermediate or meso level representation between the symbolic and the connectionist levels to be useful, its links to these levels should be clearly defined. It ought to be able to represent relations and structure other than properties. How the environment influences the conceptual system and how the symbols relate to the states need to be spelt out. The dynamics of goal-directed state changes in the space must be described.

This paper does these things. Section 2 summarises the argument for a meso level conceptual representation which was presented in Gärdenfors [15]. A comparison of alternative formulations of conceptual spaces requires them to be formally articulated. Section 3 presents definitions and results involving betweenness and distance, which provide important tools without assuming a Euclidean or even an additive structure. Section 4 applies the tools to formal definitions of the spaces and their constructs, taking guidance from findings on human conceptualisation. Section 5 introduces dynamics, which can be phrased in terms of dynamical systems, as a conceptual manipulation language using rules of association, or, in some circumstances, as manipulations of vectors. In each case, the trajectories of states can be described. These trajectories can lead to states representing complex structures and relations. Problem solving at the conceptual level reduces to micro-steps of composition and differentiation. Section 6 gives examples of conceptual spaces according to our definition. First, an example of dynamical systems is re-cast into the conceptual space framework. Then high dimensional quasi-Euclidean spaces in which generalisation is the main reasoning activity are presented, after Gärdenfors. Section 7 summarises and looks to further work and applications.

2. Gärdenfors' argument for a meso level conceptual space

This section justifies the development of a theory of representation at a level between the connectionist and the symbolic, and is drawn from Gärdenfors [15]. Also, see [13,14,16].

2.1. Symbolic representation

The crucial question Gärdenfors puts for any theory of representation is how concepts are to be modelled. On the symbolic level, basic concepts are in fact *not* modelled, rather, they are *named* by the basic symbols. Names of more complex concepts are then constructed by compositions—logical or syntactical—of the simple names. When the symbols are used for modelling logical inference, the expressions represent propositions, and a state is defined by the various logical relations among symbols. Information is processed by computing the consequences of these relations, using some set of inference rules. When symbols are used for syntactical parsing in language, strings of them are processed by different kinds of automata according to a recursive set of recursive grammatical rules. The material basis for the symbolic processes is irrelevant to the description of their results. Although symbolic sentences are assumed derived from inputs from sensory channels, the system merely performs logical operations on them.

In order to achieve its goals, a successful system is likely to have to not only form new combinations of the given symbols, but also to learn new properties from its interactions with the environment. It should be able to evolve the meaning of a concept as a result of new experiences. Changes in meaning of concepts, and development of new concepts, cannot be easily represented in formal logic systems. Concept formation techniques, such as predicate invention using inverse resolution, are based on the syntactic form of the background knowledge and observed examples: they rely on previous interpretation of the existing symbols. The underlying difficulty is the symbol grounding problem, expressed by Harnad [20] as follows: “*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?*”.

Notions of similarity and distance, which are instrumental in differentiation and identification, are difficult to model in a natural way in a symbolic system. As Quine says, “[*One*] cannot easily imagine a more familiar or fundamental notion than [similarity], or a notion more ubiquitous in its application. On this score it is like the notions of logic: like identity, negation, alternation, and the rest. And yet, strangely, there is something logically repugnant about it. For we are baffled when we try to relate the general notion of similarity significantly to logical terms” [34].

Symbolic systems are particularly vulnerable to the frame problem, that is, the specification of what in the system’s representation of the environment changes and what stays constant. This is partly because causality is not represented in first order logic. Gärdenfors also points out that the combinatorial explosion of symbolic representations of a changing world is a result of not keeping symbolic information about different domains separated.

A final problem area for symbolic systems relates to induction. If logical relations alone are used to determine which inductions are useful, the fact that all predicates are treated on an equal footing induces symmetries which are not preserved by our understanding of the inductions. “Raven” is treated on the same footing as “non-raven” in Hempel’s [21] famous example, or “green” with “grue”—the colour green till the year 2000 then blue thereafter—in Goodman’s riddle of induction [18]. While induction treats certain predicates as being

provided from the environment, the fact that the atomic predicates are taken as granted from the beginning means that much inductive processing has already been performed.

2.2. Connectionist/subconceptual representations

In connectionism, representations are via the dynamics of the patterns of activities in artificial neuron networks (ANNs), which form multi-dimensional representations. The activity of each neuron can be considered as a quality of a dimension. This way of looking at ANNs is sometimes called the state space approach. The activity of an ANN can then be represented as a vector. However, this kind of representation is, in general, different from the one studied on the conceptual level. The basic difference is that perceptually-derived information received by the receptors is tremendously rich and unstructured, and so the ANN must have complex nodal structures to transform it into a form that can be handled on the conceptual or symbolic level. Thus the state space of an ANN is of much higher dimensionality than at the conceptual level. ANNs may employ distributed representations to allow a more parsimonious structure, but this creates problems with accuracy of representation. On the conceptual level, in contrast, irrelevant information has been filtered out.

Furthermore, the metrics of the spaces on the conceptual level are in general simple in comparison to the very complex distance metrics employed by an ANN after it has been trained. In a connectionist system, distance may appear as an emergent feature, but is hard to model on a neuronal level. ANNs learn about similarities slowly and only after tailored training. Gärdenfors concedes that networks can be made more efficient by building in structural constraints when setting up the architecture of the network. However, he points out that this means that information about the relevant parts of the environment is designed into the network: the strategy presumes what we are calling the conceptual level in the very construction of the network. The representation may not be able to deal with a changing environment.

How then is the transition from the subconceptual through to the symbolic level to be made? There are systems, like ART, that perform dimensionality reduction, and which realise our notion of a conceptual space. The fundamental epistemological problem, Gärdenfors claims, is that even if the network has learned to categorise the input in the right way, it is not possible to describe what the emerging network represents because concepts are represented implicitly. There is no theory of neural networks to bridge the gap between the subconceptual and conceptual level.

2.3. Conceptual spaces

As we have said, a conceptual space is defined through a number of quality dimensions, and can for the moment be considered to be a multidimensional space (although it will be shown later to be much more). There is no unique way of choosing a dimension to represent a particular quality of the environment, even given a fixed set of perceptual input mechanisms. For example, human perception of taste appears to be generated from four distinct types of receptors: salt, sour, sweet, and bitter. The quality space representing tastes is often described as a 4-dimensional space—one such model proposed by Henning

in 1916 was of a gustatory space described as a tetrahedron—but, as with colour space, there are alternative models. An individual operates with their own set of dimensions and construction of conceptual space.

Conceptual spaces address the symbol grounding problem by giving symbols and expressions meaning via their connection to constructions in the space, such as properties and objects and the like. The form of the connections is discussed in Section 4. The assignment of meanings to the expressions on the symbolic level is therefore not arbitrary, but constrained by the underlying conceptual structure. Such a linkage is a principle of cognitive linguists [23]. Unlike a realist semantics, the resulting semantics do not presume any objects outside the cognitive structure to determine the meaning of symbols; of course, there *may* be referents external to the cognitive system.

Conceptual spaces provide a natural way of representing similarities, and relations of similarity play a crucial role in problem solving. The ability to compare things naturally is one of the major advantages of the conceptual space representation. Similarity judgements made by humans (and in many experiments, by animals) shed light on the structure of quality dimensions that individuals invoke in their conceptualisation of their environment. The similarity between two objects is agreed to be a function of (context-dependent) distance (e.g., [27]). If objects are represented as points in a conceptual space, then, roughly speaking, the similarity of two objects can be defined via the distance between the points that represent them in the space. Even abstract qualities may have a meaningful notion of distance. The phylogenetic classification of animals makes it meaningful to say that birds and reptiles are more closely related than reptiles and mammals. Relative distances can be assigned to emotional qualities, such as abhorrence as compared with hatred as compared with dislike. The minimal non-trivial structure we will require is a *betweenness* relation on dimensions, which is used to define regions and connectedness.

The theory of conceptual spaces may also indicate the direction to a solution to the frame problem. The basis is the separation into domains of the information to be represented. Finally, the question of where to direct attention is closely related to overlap of regions of interest and thus to the geometry of the conceptual space.

3. Preliminaries

This section presents notation and definitions needed to formally define properties, objects and distances, which are the foundations of conceptual spaces. The important role similarity assessment plays in human reasoning motivates the formal setting of metric spaces. Metric spaces generalise the vector spaces or orthogonal feature spaces frequently assumed in cognitive science and machine learning.

Throughout we will work within the mathematical category of *pointed metric spaces*, that is, sets with a distance function¹ d and with a distinguished point, denoted $*$, which is also known as “the point at infinity” because $d(a, *) = \infty$ for all $a \neq *$ in the space. This means that all spaces are taken to be pointed metric spaces, and all maps are taken to be

¹ That is, a bivariate, non-negative, real valued, symmetric function d such that for any x, y and z , $d(x, y) = 0$ if and only if $x = y$, and $d(x, y) \leq d(x, z) + d(y, z)$.

continuous and to map distinguished points to distinguished points. Distinguished points are a device to allow data to be “not applicable” or “not available”, as described below.

Constructing representations in terms of metric spaces rather than sets does not necessarily either constrain or add meaningful structure to them. Any set can be equipped with the discrete 0-1 metric which defines the distance between a point and itself to be zero and distances to other non-distinguished points to be 1. This trivial metric says that individual points can be distinguished, but cannot otherwise be compared: it is often implicitly applied, for example when nominal attributes are used in inductive learning algorithms. Defining conceptual spaces to be metric spaces embeds similarity comparisons into the structure, and constrains the maps used in the conceptual constructions in a sensible way. It also allows maps between conceptualisations to be defined in a sensible way, although we do not have the space to pursue these here.

A point-wise metric d can be extended to sets in a metric space through the *Hausdorff distance* d^H which, given two regions A and B , is defined by

$$d^H(A, B) = \max\{h(A, B), h(B, A)\}, \quad \text{where } h(A, B) = \sup_{a \in A} \inf_{b \in B} \{d(a, b)\}.$$

This construct is usual when going from a metric defined on points to one defined on subsets because it is a true metric, unlike the minimum separation distance which is a pseudometric when applied to intersecting sets. Use of the metric d^H will allow us to talk about the distance between properties and other constructs in conceptual space. Fig. 2 illustrates for a two dimensional Euclidean domain. In this example, the sets “tall & heavy” and “medium size” overlap: for instance, these might refer to children of a certain age, where additional factors such as bone structure alter the categorisation of size. Nevertheless, the distance between the concepts of a tall and heavy 10-year old and a medium size 10-year is intuitively not zero. The Hausdorff distance accommodates this, in this example setting the distance to be that between the lightest of the shortest “medium size” children and the lightest “tall & heavy” child.

Straight lines are fundamental to geometry. Although in Euclidean spaces the metric can be used to define lines, defining them this way is not always useful. In the following,

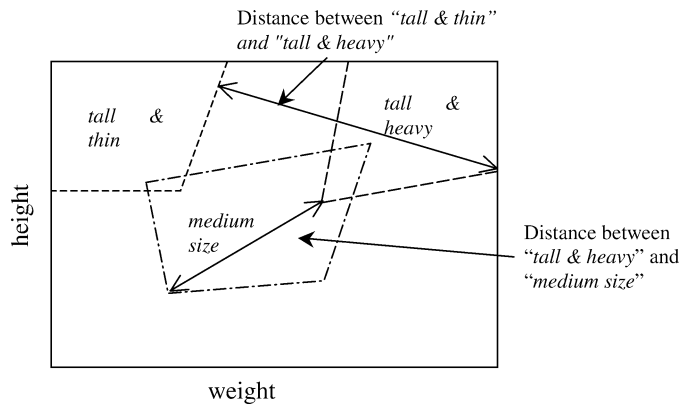


Fig. 2. Distance between connected sets in a size domain. See text.

betweenness is therefore introduced, and becomes a crucial geometrical tool, alongside distance. Betweenness is related to, but weaker than, partial order.² We nevertheless use it to define the key notions of *connectivity* and *convexity* without recourse to the metric. Connectivity and convexity in turn are used in Section 4 in defining *properties* and *prototypes*, the latter being points which, in some sense, summarise concepts. We suggest in Section 4 that properties can be formed by partitioning a space into convex sets, each containing a prespecified prototypical point. Here, we show this can be done for arbitrary metric spaces if a boundary set is allowed. To form prototypes, a sensible notion of a centroid is required even when there is no addition defined on a space. We present this at the end of Section 1. The second subsection of this section formally defines *dimensions*, *domains* and *symbols*.

3.1. Betweenness

Important structure on dimensions is provided by a betweenness relation. Given any two aspects of a representation, a third may or may not be “between” the first two. Betweenness is therefore a trivariate truth valued function. Betweenness is a conceptually ubiquitous notion, which the clinician may use in interpreting the significance of the shape of the patient’s fingernails, their pallor, or their entire demeanour. Betweenness is a weak structural requirement, for example defined naturally on graphs when a node is on a path between two other nodes.

In Euclidean space, betweenness normally takes its everyday meaning, that is, it is defined in terms of the distance metric by $B(a, b, c) \Leftrightarrow d(a, c) = d(a, b) + d(b, c)$. On the other hand, sometimes the betweenness relation defined this way is uninteresting (e.g., nothing is between anything). For example, betweenness relations are always empty when they are defined using the discrete 0–1 metric, such as might be applied to a nominal set of fruit types $\{\textit{apple}, \textit{orange}, \textit{rock melon}, \textit{pineapple}, \textit{banana}\}$. Yet complex subjective notions of taste and appearance, say, might allow some of the fruits to be related by betweenness.

Motivation for loosening the coupling of the metric and betweenness also comes from the following proposition:

Proposition 3.1.1. *Betweenness defined via metrics is not preserved under isomorphisms in the category of pointed metric spaces.*

Proof. For a contradiction, suppose a given metric d on a set D induces a non-trivial betweenness relation $B(a, b, c) \Leftrightarrow d(a, c) = d(a, b) + d(b, c)$. Define d' by $d'(a, b) = (d(a, b))^{1/2}$ for all $a, b \in D$. Then it is easy to show that d' is a metric and that the identity map between the metric spaces (D, d) and (D, d') is continuous in both directions, so that they are isomorphic. Suppose B' is the betweenness relation defined using d' . Then, for

² Given a partial order $>$ on $D - \{*\}$ and a relation B defined on $D \times D \times D$ by $B(a, b, c)$ if and only if $a > b > c$ or $c > b > a$, then it is easy to see that B satisfies parts (a) to (d) of Definition 3.1.2. However, a betweenness relation does not necessarily define a partial order: Consider a graph shaped as a “Y” with betweenness inherited from the usual ordering on a line. Given points a and c on the arms of the “Y”, b at the junction, and d at the foot, then $B(a, b, c)$, $B(a, b, d)$ and $B(c, b, d)$. It is not possible to convert these relations into expressions of the form $a < b < c$ or $a > b > c$ without arriving at a contradiction.

any distinct $a, b, c \in D$, $B'(a, b, c) = 1$ implies $(d(a, c))^{1/2} = (d(a, b))^{1/2} + (d(b, c))^{1/2}$ so that $d(a, c) = d(a, b) + d(b, c) + 2(d(a, b)d(b, c))^{1/2} > d(a, c)$ as d is a metric and $a \neq b$ and $b \neq c$. This contradiction shows that $B'(a, b, c) = 0$ and hence, B' must be trivial. \square

Borsuk and Szmielew [4] presented a set of axioms that deliver the conventional notion of betweenness on Euclidean spaces. We introduce a further two axioms, 3.1.2(e) and (f), to respectively cater for distinguished points and to provide coherence between the geometry associated with the metric and the betweenness relation. Axiom (f) ensures that betweenness is at least as rich as the metric space definition.

Definition 3.1.2 (*Betweenness*). A *betweenness relationship* B on a space D with metric d is a logical relation B on $D \times D \times D$ such that for arbitrary $a, b, c \in D$,

- (a) $B(a, b, c) \Rightarrow a \neq b, a \neq c, b \neq c$;
- (b) $B(a, b, c) \Rightarrow B(c, b, a)$; $B(a, c, b) \Rightarrow \text{not } B(c, a, b)$;
- (c) $B(a, b, c) \ \& \ B(b, c, d) \Rightarrow B(a, b, d)$;
- (d) $B(a, b, d) \ \& \ B(b, c, d) \Rightarrow B(a, b, c)$;
- (e) $\text{not } B(a, b, *) \ \& \ \text{not } B(a, *, b) \ \& \ \text{not } B(*, a, b)$ for all $a, b \in D$;
- (f) $d(a, c) = d(a, b) + d(b, c) \Rightarrow B(a, b, c)$.

The set of points x satisfying $B(a, x, b)$ can be said to form a line or a path between a and b . The second condition in (b) ensures there are no cycles, such as the circle with its usual local notion of betweenness. However, there can be multiple paths between two points.

Connectivity and convexity are important descriptors of the natural environment, because primitive entities are almost always perceived to be spatially connected, and, often, to be convex, while primitive events are seen to be temporally connected and hence convex in uni-dimensional time. Connectivity and convexity are therefore important in conceptual space constructions. The next definition defines convexity in terms of betweenness, and defines regions as connected sets. Connectivity is defined in a non-standard way using the betweenness relation that allows some finite sets to be connected.³

Definition 3.1.3 (*Regions, connectivity and convexity*).

- (a) A space A is called r -convex if it is a singleton or if for any pair $a = c_0, b = c_r \in A$, there exist $r - 1$ elements $c_1, c_2, \dots, c_{r-1} \in A$ such that $B(c_i, x, c_{i+1}) \Rightarrow x \in A$. A 1-convex space is called *convex*. It is easy to see that if A and A' are convex, then so is $A \cap A'$.
- (b) The *convex closure* of the set A is the smallest convex set containing A . The construction of such a set is described in Definition 3.1.5.

³ A connected set A is usually defined to be one in which, given arbitrary a, b in A , there is a continuous map f from the unit interval into A such that $f(0) = a$ and $f(1) = b$. Definition 3.1.3 excludes some sets which are connected under the usual definition, such as curved lines in Euclidean 2-space. However, these are limit points of sequences of connected sets. That is, if A is connected according to the usual definition, there exist A_n , connected under 3.1.3, such that $\lim_{n \rightarrow \infty} \sup\{\inf\{d(a, b) : a \in A_n, b \in A\}\} = 0$.

- (c) A subset A of a space D with a betweenness relation B is said to be *connected* if
- (i) it is a singleton, or
 - (ii) for any pair $a, b \in A$, for some $n > 0$ there is a sequence $c_0 = a, c_1, c_2, \dots, c_{n-1}, c_n = b \in A$ such that $B(c_i, x, c_{i+1}) \Rightarrow x \in A$.
- (d) A *region* of D is a connected closed subspace of D .

Under its usual definition of convexity, any Euclidean space can be divided into a set of n convex regions, disjoint except at boundaries, where the i th region is chosen to contain a pre-specified point x_i . This is done in a process called *Voronoi Tessellation*, which assigns points in the space to the i th region if and only if x_i is the closest of the pre-specified points (see, for example, [28]). Without distance, such a neat construction is not possible. However our definition of convexity using betweenness needs to support a similar tessellation in any domain, which will partition a Euclidean space into conventionally convex sets if the usual notion of betweenness applies.

Standard Voronoi tessellation involves boundary points which are assigned to more than one of the convex sets. In this case, boundaries are hyperplanes in the Euclidean space. The general boundary between convex sets is more interesting, and can support indeterminacy in assigning points to regions. To motivate the definition of a boundary, we first present an example. Then we prove that a tessellation can always be constructed so that pre-specified points belong to mutually disjoint convex regions which cover the space, apart from a boundary set.

Example 3.1.4. There is a set with a betweenness relation which cannot be partitioned into two non-empty convex sets.

Let $S = \{Anne, Belinda, Carol, Bob, Tom, T0, T1, T2, T3\}$ have the following betweenness relations:

$$B(Belinda, Bob, Carol), B(Belinda, Tom, Carol), B(Bob, Anne, Tom) \quad (3.1)$$

$$B(Carol, T0, Anne), B(Carol, T1, Anne), B(T0, Belinda, T1) \quad (3.2)$$

$$B(Anne, T2, Belinda), B(Anne, T3, Belinda), B(T2, Carol, T3) \quad (3.3)$$

It is straightforward to check that these relations satisfy the definition of a betweenness relation. (The betweenness relation might refer to relative “reliability”; *Anne, Belinda, Carol, Bob* and *Tom* might be general practitioners, and *T0, T1, T2, T3* might be medical tests.)

The proof that S cannot be partitioned as described starts by assuming that S is partitioned into two sets, and shows that if convexity is preserved, all the elements of the set are forced to belong to one set. That is, the other set is empty and the partition is trivial. Name the two sets *Reliable* and *Unreliable* respectively. Two of the elements *Anne, Belinda, Carol* must belong to the same set, say *Belinda, Carol* are *Reliable*. Now we show that all other elements are *Reliable*. By convexity, $Belinda, Carol \in Reliable \Rightarrow Bob, Tom \in Reliable$ from (3.1). Then $Bob, Tom \in Reliable \Rightarrow Anne \in Reliable$ from (3.1). Now that *Anne, Belinda* and *Carol* are all *Reliable* it follows from (3.2) and (3.3) that the remaining elements *T0, T1, T2, T3* are also. Clearly, the same argument follows if we started by assuming that *Anne* and *Belinda* are *Reliable* or *Carol* and *Anne* are *Reliable*.

So, if the set S is to be partitioned into 2 convex sets because *Anne* is *Unreliable* and *Belinda* is *Reliable*, what can be done? Suppose the element *Carol* is judged *Reliable*. Then in order to maintain convexity, *Tom* and *Bob* must be added to this set, and hence *Anne* would have to be added—but she is *Unreliable*. Similarly, *Carol* cannot be added in with *Anne*. So *Carol* must be a *boundary point*. This suggests the nature of the boundary of a partition of sets, as those points which cannot be included in the convex sets without causing a violation of convexity. Points in the boundary can be grouped according to the betweenness relations they have, to give the boundary a structure we do not explore further here. Boundary points are fuzzy in the sense that they can be viewed as belonging to more than one grouping, which might occur, for example, if *Carol* is judged as reliable in comparisons involving doctors and tests, but is unreliable in comparisons involving doctors.

Definition 3.1.5. Assume a partition of the set A into sets E, A_1, \dots, A_N . Then the set E is a *boundary* of the sets A_i if for every element $x \in E$ there is not a set A_i for which the convex closure of $A_i \cup \{x\}$ remains disjoint from $\bigcup_{i \neq j} A_j$.

Before going on to our main theorem, we present a construction of the convex closure of an arbitrary set which will be used in the proof. The theorem proof will be by construction, assuming the Axiom of Choice and using transfinite induction, ie. induction past the usual countable infinite sequence [43].⁴

Construction 3.1.6 (*of the convex closure of an arbitrary set*). Given the set S , we construct a convex set S' such that $S \subseteq S'$ and for every convex set S'' , if $S \subseteq S''$ then $S' \subseteq S''$. We do this by inductively constructing sets S_n such that $S \subseteq S_n$ and $\bigcup\{S_n: n \geq 0\}$ is convex. Put $S_0 = S$ and for each $n \geq 0$ put $S_{n+1} = \{x: x \in S_n \text{ or } B(s, x, t) \text{ for some } s, t \in S_n\}$. Put $S' = \bigcup\{S_n: n \geq 0\}$. We show S' is convex. Let $a, b \in S'$ and suppose there is x such that $B(a, x, b)$. Each element a, b must appear in S_0 or one of the S_{n+1} constructions, after which they remain. Choose k to be the maximum of these two numbers, that is, $a, b \in S_k$. Then $x \in S_{k+1}$ and so $x \in S'$. It remains to show that S' is the smallest such set, that is, $S' \subseteq S''$. This follows from the observation that each $S_n \subseteq S''$ since at every n all the elements of S_n are necessary for S'' to be convex.

To gain insight into the following theorem, consider the case of a set $S = size$ in Euclidean 2-space, with, say, 3 points which have been assessed as respectively being *small*, *medium* and *large*. Fig. 3 illustrates. To partition the set into 3 disjoint convex sets with boundaries requires defining 3 straight lines across the set which separate the points. In a Voronoi Tessellation this is done by defining the boundaries to be the points which are equi-distant from at least two of the pre-specified points. This construction, however, does not produce convex sets in general non-Euclidean spaces in which the betweenness and the distance function are de-coupled. In the construction of the proof below, therefore, convex

⁴ Although transfinite techniques are common in mathematical logic, they are unusual in AI and cognitive science because, although constructive, they are not computable except on finite sets. However, the aim of the theorem is to guarantee the existence of such partitions, not to construct them.

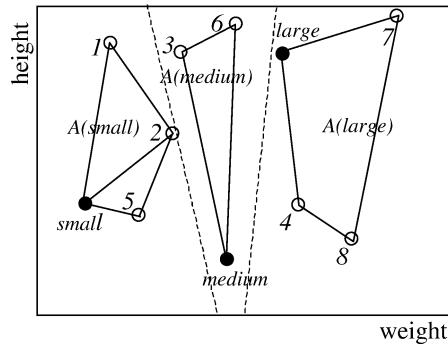


Fig. 3. Example of the tessellation of a space about prescribed points. Three convex sets $A(\text{small})$, $A(\text{medium})$ and $A(\text{large})$ are grown about pre-designated points small , medium and large , shown as the centres of the solid circles. The numbered points indicate the order in which the points 1–8 have been considered. For a point to be added to a convex set, the convex closure must not intersect the other convex sets. The points 1–3 can be added to arbitrary sets $A(x)$. In this example, points 1 and 2 are added to $A(\text{small})$, and point 3 to $A(\text{medium})$. Point 4 cannot be added to $A(\text{small})$ without the convex closure intersecting the line $\text{medium}-3$; so in this example it is added to $A(\text{large})$ although it could instead have been added to $A(\text{medium})$. Point 5 can only be added to $A(\text{small})$, and so on. The dotted lines indicate the boundaries of the final tessellation. The position of the boundaries depends on the (transfinite) ordering of the points in S . See proof of Theorem 3.1.7.

sets are “grown” about the pre-specified points by iteratively selecting an arbitrary point and an arbitrary convex set and, if possible, adding all the points in the convex closure of the selected set and point. When this cannot be done without intersecting some other convex set, another set is selected. If no set can be grown while maintaining convexity, the point is instead put into the boundary.

Theorem 3.1.7. *Given $n > 0$, any domain S can be divided into n disjoint convex regions plus a boundary set, which can be chosen so that the i th convex region contains a pre-specified point x_i .*

Proof. Let $\{s_\alpha : \alpha < |S|\}$ be a listing of the points in the domain S , indexed by the ordinals α . (This may be a transfinite set in the case that the domain is uncountable, for example, is the reals; such a listing is still possible by the Axiom of Choice.) For each of the elements x_i we construct a set A_i such that $x_i \in A_i$ and the sets are mutually disjoint.

Put $E = \emptyset$ and $A_i = \{x_i\}$ for each $i \leq n$. Let $S_0 = \bigcup_i \{A_i\} \cup E$ and for each $\alpha < |S|$, construct $S_{\alpha+1}$ as follows. If s_α is not already in S_α arbitrarily choose a set A_i from S_α with the property that the convex closure of $A_i \cup \{s_\alpha\}$ does not intersect any of the other sets A_j or E (if there is such a set A_i). Rename the convex closure of $A_i \cup \{s_\alpha\}$ as A_i . If there is no such A_i then add s_α to E and rename E . Form $S_{\alpha+1}$ from the union of the new sets as before. Clearly $S_\alpha \subseteq S_{\alpha+1}$, each of the new sets E and A_i is convex, and all the new sets are mutually disjoint.

It remains to perform the construction for S_α when α is a limit ordinal, that is, an infinite union of predecessor ordinals. Put $S_\alpha = \bigcup \{S_\beta : \beta < \alpha\}$. Define the new sets E and A_i to be the union of the respective earlier sets. Now convexity follows using the same argument

as in the convex closure construction, and similarly for mutual disjointness since if there were a mutual element between two sets there must be a $\beta < \alpha$ where this occurred.

In particular, this argument holds for the limit ordinal $\alpha = |S|$, so it remains to show that $S_{|S|} = S$. But this follows from the fact that every element of S has been added to one of the sets since the construction followed the enumeration $\{s_\alpha: \alpha < |S|\}$. Finally, we observe that the set E in $S_{|S|}$ is a boundary set since every element added to a previous set E was the result of the fact that the element could not be added to a set A_i and still maintain convexity. \square

The ability to partition spaces into convex sets around prototypes can be assumed in the remainder of this paper because of this important theorem.

The final result in this section concerns the centroid of a set of points or of a subspace of an arbitrary metric space. The space cannot be assumed to be a vector space with addition and scalar multiplication. So instead, the centroid is defined in terms of the distance function, which takes points in the space to the reals, where addition is defined. The definition reduces to the usual ones when the space is Euclidean. The definitions are, as usual, with respect to compact⁵ spaces, which in Euclidean spaces are the closed bounded subsets, and in finite spaces are arbitrary subsets.

Proposition 3.1.8.

- (a) Suppose $v_1, \dots, v_n \in A$ for some compact space A with metric d , and $h_n(y) = \sum_{i=1,n} w_i d(y, v_i)^2$ for some positive weights w_i . Then
 - (i) h_n attains its minimum on A and
 - (ii) if A is a convex set in a Euclidean vector space, then h_n attains its minimum on the weighted average of the points v_i .
- (b) Suppose w is a non-negative weighting function on a compact space A . Then a centroid of A can be defined which reduces to the definition (a) when A is finite, and has the usual meaning when A is Euclidean and convex.

Proof.

(a) A minimum exists because $h_n = \sum_{i=1,n} w_i d(-, v_i)^2$ is a continuous function from A into the real line, and so the compact set A has closed and bounded image $h_n(A)$ in the reals. If z is the minimum, call z the centroid of the points v_i . If A is Euclidean m -space, suppose $v_i = (v_{i1}, v_{i2}, \dots, v_{im})$ and $x = (x_1, x_2, \dots, x_m)$. Then

$$h_n(x) = \sum_i w_i |x - v_i|^2 = \sum_{j=1,\dots,m} \sum_i w_i (x_j - v_{ij})^2.$$

This attains its minimum where $\partial h_n(x)/\partial x = 0$. This is exactly when $x_j = \sum_i w_i v_{ij} / \sum_{i=1,n} w_i$ for each j , that is, when $x = \sum_{i=1,n} w_i v_i / \sum_{i=1,n} w_i$, which is the weighted average of the points. If A is convex, $\sum_{i=1,n} w_i v_i / \sum_{i=1,n} w_i$ is in A so the weighted average is the centroid z .

⁵ A compact set A in a metric space is one in which each infinite subset has a limit point. It is closed and bounded, that is, for any $\varepsilon > 0$, A is the union of a finite number of subsets of diameter ε [40].

(b) Compactness of A means a sequence of sequences $S_n = \{v_{n,1}, \dots, v_{n,N(n)}\}$ can be defined such that, for arbitrary $x \in A$, $\min_{i \leq N(n)} d(x, v_{n,i}) < 1/n$ and S_n is a shortest sequence satisfying this condition. Thus for each n , $A = \bigcup_{k>0} \{x \in A: d(v_{n,k}, x) < 1/n\}$. Let z_n be the centroid of the points in S_n (that is, a point minimising $h_n(y) = \sum_{i=1, N(n)} w(v_{n,i}) d(y, v_{n,i})^2$). By compactness again, the sequence of centroids $\{z_n: n > 1\}$ has a limit point. Define the centroid z to be a limit point.

Suppose A is Euclidean, and the sets S_n have been defined as above. As $n \rightarrow \infty$, A can be partitioned into $N(n)$ disjoint sets on which g and x become approximately constant, which each contain a distinct point $v_{n,i}$, and which vary in size by an arbitrarily small amount. Hence

$$\left| \int_A w(x)x \, dx \Big/ \int_A w(x) \, dx - \sum_{i=1, \dots, N(n)} w(v_{n,i})v_{n,i} \Big/ \sum_{i=1, \dots, N(n)} w(v_{n,i}) \right| \rightarrow 0$$

as $n \rightarrow \infty$.

Now $\sum_{i=1, \dots, N(n)} w(v_{n,i})v_{n,i} / \sum_{i=1, \dots, N(n)} w(v_{n,i})$ minimises $\sum_{i=1, \dots, N(n)} w(v_{n,i})|y - v_{n,i}|^2$. Hence if $z' = \int_A w(x)x \, dx / \int_A w(x) \, dx$ then $z' = \lim_n \{z_n: z_n \text{ minimises } \sum_{i=1, \dots, N(n)} w(v_{n,i})d(y, v_{n,i})^2\}$. Thus the centroid has the usual definition. \square

3.2. Dimensions, domains and symbols

A central notion of conceptual spaces is that of a quality dimension, used to represent the qualities of objects and concepts. The very fact that two aspects of a representation have been assigned to the same dimension means that they are in some sense alike. When a conceptual space is used as a framework for scientific theory or for construction of an artificial cognitive system, as in a robot or an automated diagnosis system, the structure of dimensions is chosen by the scientist or system builder. The choice of dimensions will depend heavily on the underlying theory and on what environmental sensors can be utilised by the system—range sensors for robots or microscopic instrumentation for analysing tissue, for example. In contrast, the dimensions of a conceptual space developed through self structuring cannot readily be obtained from either the perceptions, or the actions, of the animal or artificial system. They have to be inferred. The best known psycho-experimental method for doing this is multi-dimensional scaling. This starts from a subject's judgements—either verbal or implied—about the similarity of pairs or sets of stimuli as the stimuli are varied along a number of potential dimensions. The minimum number of dimensions required to adequately explain the experimental data is calculated, although even when this is done, the psychological interpretation of the dimensions generated by the algorithm may not be obvious [8].

Dimensions may provide only a partial handle on the conceptual space used in cognitive processing. The following definition captures the sometimes loose tie between the space and its dimensional structure (discussed again later). This tie is achieved by associating the space to a dimension only through a projective mapping, which must, however, preserve betweenness, as this is the primary ingredient of geometry in our spaces. Notions of complete and covered dimension sets are introduced as nomenclature for cases in which the conceptual space is fully described by the dimensions.

A serious problem with describing an object as “a point in conceptual space” is that, often, dimensions in the space are irrelevant or inappropriate to a particular object or concept. Furthermore, an object’s values on relevant dimensions may be unknown to the system. As indicated above, the distinguished element $*$ has been introduced as a technical device to allow for these cases.

Definition 3.2.1 (*Dimensions, completeness and coverings*).

- (a) A *dimension* of a space D equipped with a betweenness relation B is a space D_i equipped with
 - (i) a metric d_i ;
 - (ii) a betweenness relation B_i ;
 - (iii) a continuous projection: $\pi_i D \rightarrow D_i$ such that π_i preserves betweenness, i.e., given $a, b, c \in D$ such that $B(a, b, c)$ and the elements map to distinct elements in $D_i - \{*\}$, then $B_i(\pi_i(a), \pi_i(b), \pi_i(c))$.
- (b) A set of dimensions $\{D_i, i = 1, \dots, N\}$ is *complete* for D if, for any $x, y \in D$ with $\pi_i(x) = \pi_i(y)$ for every $i \leq N$, then $x = y$. Thus for example, the Red and Green colour dimensions, without Blue, are not a complete set of dimensions for colour space.
- (c) A set of dimensions $\{D_i, i = 1, \dots, N\}$ is *covered* by D if for every member p of $D_1 \times D_2 \times \dots \times D_N$ there is an element $x \in D$ such that $p = \{\pi_1(x), \pi_2(x), \dots, \pi_N(x)\}$. The NCS colour spindle [19] is an example of a product of dimensions which is not covered: there are points on one colour axis which are projected upon only when the other projections fall in a certain range.

If a set of covered dimensions $\{D_i, i = 1, \dots, N\}$ is complete for D , any element in D can be represented as a *unique* N -tuple $\{y_1, y_2, y_3, \dots, y_N\}$.

Gärdenfors [15] identifies conceptual spaces with spaces generated by a set of dimensions. Our more general definition allows the possibility of learning through discovering more structure. In the physical environment there are relationships—sometimes subtle and sometimes perceptually evident—between qualities. An example of a subtle relationship is the mass, volume and pressure of a gas, of a perceptually evident one is the hue, saturation and intensity of colour. In human conceptualisation there are therefore dimensions which are inextricably related, in the sense that one cannot assign an object a value on one dimension without giving it a value on another. These dimensions are called *integral* in the cognition literature, and include the pitch and volume of sound. Dimensions which are not integral are said to be *separable*.

Children learn to separate dimensions as part of their development, the classic example being confusion amongst volume and spatial dimensions. More generally, training and experience allow people to separate dimensions in stimuli they initially find integral. In animal and human conceptualisation, separability of dimensions is defined through the form of the distance metric that appears to be used in judgements involving multiple dimensions: if a city block metric is not fitted by experimental data on conceptual distances involving multiple dimensions, then the dimensions are assumed to involve at least one pair of integral dimensions (e.g., [39]). The metric structure of psychological space has been the subject of extensive experimental study [27].

Despite difficulties in interpreting experimental results, *formally* distinguishing integral and separable dimensions is straightforward. Part (a) of the following definition merely groups those dimensions which are integral into what will be called domains. Part (b) allows that not all points in the product of a set of integral dimensions may be conceptually feasible, as is the case with colour as a three dimensional domain. After this definition, we tend to talk in terms of domains, as the molecular units which provide the stuff of concepts, but which are themselves composed of dimensions at the atomic level. Mostly, we won't look into how the domains are structured—how the atoms are bound—and so the maps ϕ_i in this definition are rarely referred to in later definitions. However, they are an important part of the conceptual structure, and, as the psychological literature indicates, affect similarity judgements.

Definition 3.2.2 (*Integral and separable dimensions, and domains*).

- (a) Consider a space D with dimensions $D_i, i = 1, 2, 3 \dots$. Suppose that for some M less than or equal to the number of dimensions, the set $\{1, 2, \dots, M\}$ is partitioned by M' non-singleton subsets Y_i (i.e., each Y_i has more than one element, and $Y_i \cap Y_j = \emptyset$ for each $i \neq j$, and $\bigcup\{Y_i\}_{i=1, \dots, M'} = \{1, 2, \dots, M\}$). For $i > M'$, set Y_i to be the singleton set containing the integer $M - M' + i$.

Then call a dimension *integral* if it indexed by Y_i for $i \leq M'$, and say that any dimensions indexed by the same set Y_i are *integrally related*.

Any two dimensions which are not integrally related are called *separable*.

- (b) For $i > M'$, let $\Delta_i = D_{i+M-M'}$.

For $i \leq M'$, let Δ_i be isomorphic to $\pi(D)$, where $\pi : D \rightarrow \prod_j \{D_j : j \in Y_i\}$ is induced by the projections $\pi_i : D \rightarrow D_i$ and there is a set function $\phi_i : \prod_j \{D_j : j \in Y_i\} \rightarrow \Delta_i$ which is an isomorphism of metric spaces on $\pi(D)$, and is $\phi_i(x) = *$ for $x \notin \pi(D)$.⁶

Abuse notation by letting $\pi_i : D \rightarrow \Delta_i$ be the composite $\phi_i \pi_i$. Similarly abuse notation by denoting the induced metric on Δ_i by d_i and the induced betweenness relation by B_i .

The spaces Δ_i are called the *domains* of D determined by the *dimension partition* $\{Y_i\}$ and are called *integral domains* if they are formed from more than one dimension. The map π_i is called the projection onto the i th domain.

A set of domains can be complete for and/or covered by D in analogous definitions to 3.2.1. A product of domains, each with a distance measure, is equipped with a distance measure through the city block metric.⁷ This is a consequence of the definition of domains: the dimensions in different domains are independent. So distances should be

⁶ Here and elsewhere, if A_i is a space with a distinguished point $*$, a product $A_1 \times A_2 \times \dots \times A_n \equiv \prod_{i=1, \dots, n} A_i$ is understood to mean the set $\{(a_1, a_2, \dots, a_n) : a_i \in A_i\}$ with distinguished point $(*, *, \dots, *)$. The product takes the city block metric (footnote 7) except on points such as $(*, 1, 1, \dots)$, for which the behaviour of the metric is discussed in Section 4. Given a set of maps $\beta_i : B \rightarrow A_i$, $\prod_{i=1, \dots, n} \beta_i : B \rightarrow \prod_{i=1, \dots, n} A_i$ is the map defined componentwise.

⁷ The city block metric d on a product of metric spaces $A_1 \times \dots \times A_n$ with respective metrics d_i is given by $d(\mathbf{x}, \mathbf{y}) = \sum_i d_i(x_i, y_i)$, for $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$. In general, a Minkowski metric is defined as $(\sum_i d_i(x_i, y_i)^r)^{1/r}$, which if $r = 2$ is called a Euclidean metric and if $r = \infty$ is called a chess board metric.

additive across domains. In contrast, even though Definition 3.2.2(b) makes a domain Δ_i isomorphic to a space equipped with a city block metric composed from distances in the dimensions of Δ_i , Δ_i itself might have the Euclidean or some other distance.

The final definition of this section distinguishes a set of dimensions, the *symbol subspace*. As discussed earlier, developing the links that perceptual and symbolic systems have with a conceptual system is a key contribution of this paper, and the symbolic communication is through this subspace.

A modelling or construction device used in almost all cognitive systems is to link the *strength* of activity on a symbolic dimension with the pattern of activity in subsymbolic dimensions (e.g., in fuzzy pattern recognition systems, ART networks etc.). In the next sections, we formally explore in conceptual space terminology what happens in systems such as ART or PATON [29] which allow the external activation of a symbol dimension to change the current value on non-symbol dimensions and vice versa. Therefore the explicit link to magnitude in the following definition of a symbol dimension formalises what is routinely assumed by modellers of cognitive representations.

Definition 3.2.3 (*Symbol subspace*). A *symbol subspace* of size n is a space generated by n dimensions which are each isometric⁸ to $\{*\} \cup (0, 1]$. A dimension in the symbol subspace is called a *symbol dimension*.

If the distinguished point $*$ is called 0 then, as sets, a symbol dimension can be identified with $[0, 1]$.

4. Conceptual spaces and their key constructs

This section presents a generic formulation of conceptual spaces which attempts to constrain the representation as little as possible, while capturing the essential features of conceptual spaces teased out in [15]. These are the notion of dimensions and domains, of properties as regions in domains, and of concepts defined via properties. Selective associations between such regions and symbols are the basis of representation and communication.

The first section presents a specification of conceptual spaces, setting out the conditions for a set to qualify as a conceptual space. The dimensions are the “handles” used to describe the base conceptual space. The key constructs of objects, properties and concepts, together with generalised distance functions, are then defined formally. These give conceptual spaces the power of geometric representation. The final section describes an *individual’s conceptualisation* as a subspace of conceptual space, in which symbols are linked to regions and to prototypical values in concept space. This allows properties to be described in terms of how an individual conceives that a domain is associated with a symbol. The richer structure necessary for reasoning is dynamically derived from these associations, as shown in Section 5.

⁸ An isometry between spaces A and A' with distance metrics d and d' respectively is an isomorphism φ such that $d(x, y) = d'(\varphi(x), \varphi(y))$ for all $x, y \in A$.

4.1. Conceptual space definition

We are almost in a position to specify a conceptual space: We have a way to say that things are not applicable or not known; we have a way of identifying concepts, properties and objects, as regions or points, and of naming them through symbols—so we can label concepts as *dog* or *chair* etc.—; we have a way of representing how applicable each such symbol is to a conceptual state through the activity on the symbol dimension; and we have a way of allowing for domains like colour which are not simply products of dimensions.

One remaining complication is the composition of concepts, such as “a *woman* in a *yellow dress* in a *yellow Volkswagen*” and “*pale* blood cells and *pallid* patient”. These involve shared concepts, such as “yellow” in the first example or “pale” in the second, which need to be bound to the correct object or concept. The *binding problem* is a recurring representational theme, for which the mechanisms used in the brain are still not known (e.g., [30]). Activation related to a single concept may, however, be identified by carrier frequency or phase in a multiplexed signal [12]. We employ an equivalent solution for the binding problem, viz. multiple copies of a *base conceptual space* are made, with the distinguished point indicating when a copy is not involved in a complex concept.⁹

A final complication is that the representation arguably ought to include any constructions used to support the dynamics in conceptual space, just as human working memory is part of memory. The key construction is the “spotlighting” of regions in conceptual space relevant to the current problem-solving task. These regions are specified in a buffer, whose length is included as a parameter in the following definition of the conceptual space. The form and use of the buffer is described fully in Section 5.

Definition 4.1.1 (*Conceptual space*). A conceptual space is defined by the specification of

- A space C , called the base conceptual space, with distance metric d , and a betweenness relation B .
- A decomposition of C as the product $C = D_L \times D$ such that d is the city block metric formed from the metrics on D_L and D . D_L is called the *symbol space* and is isometric to a product of closed unit intervals $[0, 1]^l$ for some $l > 0$, as described in 3.2.3.
- A finite index set Dim together with a partition $\{Y_i\}_{i \in Dom}$ of Dim into subsets indexed by $Dom = Dom(int) \cup Dom(sep)$, in which $i \in Dom(sep)$ if and only if Y_i has only one element.
- The association of each $i \in Dim$ with a space D_i called a *dimension* of D , defined as in 3.2.1(a), and the association of each $i \in Dom$ with a space Δ_i called a *domain* of D , defined as in 3.2.2.
- An integer $v > 0$ which is the number of copies of C in the conceptual space; these copies are referred to as *levels* in conceptual space. The product C^v is equipped with a distance measure using the city block metric.
- A set $(C[C])^s$ called the *attention buffer*, where $s > 0$ and $C[A]$ denotes the set of all connected subsets of A .

⁹ Chella et al. [9] take points in conceptual space to be a sequence of superquadric surfaces, and complex objects to be sets of such points with a symbolic layer describing various types of motion.

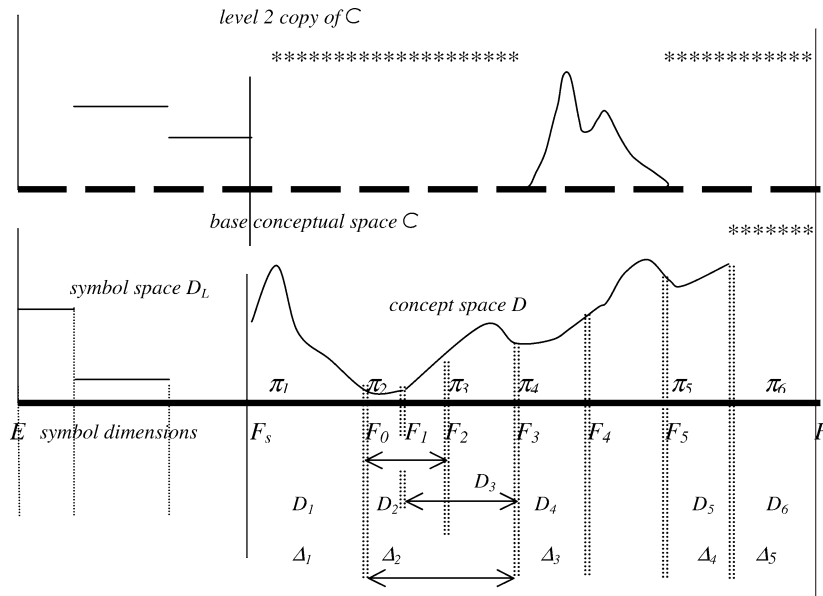


Fig. 4. Illustration of defining structures in conceptual space. The conceptual space has levels, each structured as the base conceptual space C which, in this example, is supposed to be a metric space of piecewise continuous functions into $[0, 1] \cup \{*\}$. The base space is the product of a symbol space D_L and a concept space D , which project under maps π_i onto dimension spaces, in this example, by restriction onto closed intervals in $[E, F]$. Symbol space is isometric to a product of intervals $(0, 1]$, each with distinguished point $\{0\}$. Products of some dimensions map onto integral domains Δ_k . In this example, domains are formed whenever the intervals which define dimensions overlap, and points in a domain space are the union of functions defined on such dimensions which agree on the overlap. See text.

Recall that spaces and maps are defined in the category of metric spaces with distinguished points, so that all maps in this definition are continuous and take distinguished points to distinguished points.

Fig 4. illustrates the main aspects of the definition, using a family of waveforms to represent conceptual space. Specifically, conceptual space is a family of functions defined on a finite interval $[E, F]$, taking values in $[0, 1] \cup \{*\}$. A dimension is the restriction of the family to a closed subinterval of $[E, F]$. Multi-dimensional domains are formed when intervals defining dimensions overlap. Symbol dimensions are defined on a subinterval, $[E, F_s]$ say. Functions in the family are constants in $[0, 1]$ on symbol dimensions. They are continuous in concept space except when taking the value $*$. The distance between two functions f and g is the average absolute difference of functions over the relevant interval I , that is, $|I|^{-1} \int_I |f(x) - g(x)| dx$. Betweenness is derived from betweenness of the values of functions, that is $B(f, g, h)$ if and only for all $x \in I$ either (a) $f(x) < g(x) < h(x)$ or $f(x) = * = g(x) = h(x)$ or (b) $f(x) > g(x) > h(x)$ or $f(x) = * = g(x) = h(x)$.

Another example of a conceptual space is a 2-layer network, in which case nodes are dimensions, relative magnitude of activation at a node gives distance in that dimension, the

upper level nodes form symbol space, and associations between nodes at the lower level lead to non-trivial domains.

A final example of conceptual spaces as defined by 4.1.1 is a set of Euclidean spaces, in which each component takes values in the reals, the integers or a finite ordered set, each with infinity adjoined. Betweenness has the conventional definition. Such spaces subsume many of the feature spaces used in practical and theoretical work on classification. Note, however, that our definition explicitly recognises the need to name classes and feature dimensions in the symbol space D_L .

Just as the language of a logic or the structure of a network constrains what can be done, the conceptual space specification constrains the possible conceptual richness of representation. The ability to represent the environment can grow through a variety of mechanisms which reflect human development. Specifically, aspects of the development of an individual's ability to conceptualise, as opposed to development of what they know or believe, can be modelled within the framework of Definition 4.1.1, through:

- (i) Enlarging the base conceptual space to incorporate experience; for example, enlarging the family of functions in Fig. 4.
- (ii) Learning new dimensions, and the way to project onto them; for example, forming a dimension from the restriction of functions to the interval $[F_4, F_5]$ in Fig. 4.
- (iii) Acquiring new names (increasing the dimensionality of symbol space); for example, splitting a symbol dimension in Fig. 4.
- (iv) Unravelling dimensions in domains (moving dimensions from the set of integral domains to the set of separable domains), or vice versa; for example, learning domain Δ_2 in Fig. 4 to be separable dimensions of restrictions of functions to $[F_0, F_1], [F_1, F_2], [F_2, F_3]$.
- (v) Being able to formulate more complex concepts by increasing the number of levels of composition (that is, the parameter v in 4.1.1); or
- (vi) Increasing the ability to focus on areas (by increasing the parameter s in 4.1.1).

The definition of conceptual space can be enhanced to increase the degree to which the space is constrained by its dimensions. For example, a *completely dimensioned* conceptual space could be defined as one in which the dimensions are complete. As well, the spaces and maps in the definition could be specified as belonging to a category within the category of pointed metric spaces, eg. pointed Euclidean spaces. The distance metrics could also be constrained.

4.2. Objects, properties and concepts

We can now describe the key constructs of objects, properties, and concepts. Properties are the building block of representations. Many variants are required to capture the subtleties exhibited in different knowledge domains. We take as a basic property what Gärdenfors [15] terms a *natural property*, one that is defined on a convex region. Because of Theorem 3.1.7, we know that any domain can be partitioned or tessellated into a set of convex regions and a boundary set. The following definition also allows *graded membership* of points in the regions which define properties, so that for example some "black" colours can be "less black" than others; these gradings might be applied to members of the boundary sets in a tessellation of the domain. There are many other useful

definitions that might be applied to subcategorise properties and concepts. For example, *fuzziness* could be defined, as could *action properties* which have temporal duration, to capture properties like “fast heart rate”.

Definition 4.2.1.

- (a) A *simple object* is defined to be a point in base conceptual space C . A *complex object* is a finite union of simple objects, that is, a set of points at different levels in the conceptual space.
- (b) A (natural) *property* is defined to be a convex region in a domain, and a *complex property* is a finite union of natural properties.
 - A property P on domain Δ is *graded* if there is a *grading* function $g : \Delta \rightarrow [0, 1]$ such that $g(x) > 0$ if and only if $x \in P$ and $B(a, x, b) \ \& \ a, b \in P \Rightarrow g(x) \geq \min(g(a), g(b))$.
- (c) A *concept* $c(P_1, P_2, \dots, P_n)$ involving complex or natural properties P_i with domain $\Delta_{j(i)}$ is a region in D such that for arbitrary $x \in c(P_1, P_2, \dots, P_n)$ and for each $i \leq n$, $\pi_{j(i)}(x) \in P_i$ and for each domain $k \notin \{j(1), \dots, j(n)\}$, $\pi_k(x) = *$. So $c(P_1, P_2, \dots, P_n)$ is contained in $\bigcap_i \pi_{j(i)}^{-1}(P_i)$, the intersection in D of the inverse images of the properties defining the concept.

Note that concepts are therefore regions in *conceptual space*, whereas properties are regions in *domains*.¹⁰ Thus in general, a property is not a concept. In the example of Fig. 4 a concept might involve a subset of functions defined over the entire interval between and including D_4 and D_5 , which is more complex than the union of functions projecting onto those dimensions. The reason for representing concepts this way is to allow them to be more than the dimensions that are used to describe them. Thus the concept of “dingo” may be more to an individual who fears or loves the animals than the properties that they can list, whereas to someone who does not know what a dingo is, a listing of the properties may equate to their concept “dingo”. The clinician can list the patient’s self reported condition, their own reading of the signs, and the laboratory reports, and still have not captured the conceptualisation on which they will make a diagnosis.

Note, also, that under our definition, a concept maps into the product of sets of properties in a product of domains $\Delta_1 \times \Delta_2 \times \dots \times \Delta_n$, but that not all points in a product of properties may be in the concept. That is, the concept does not necessarily cover the product of its properties. Thus in the depiction in Fig. 2, if the projection of the concept *medium* is onto intervals *medium weight* and *medium height* the concept does not cover points of the product *medium weight* \times *medium height* which are at the high end of *medium weight* and the low end of *medium height*.

Not all qualities are equally important in defining a concept. Thus red blood cell volume is the defining and hence most important quality, for the concept “anaemia”, drawing on scientifically measured rather than perceptual domains. The importance of a domain to a concept may be inversely related to the relative size of properties, since if a domain is less important, then values in it may be relatively arbitrary. In any case, the need to establish the relevance of qualities or attributes to categorisation tasks is well understood. The following

¹⁰ If a space is complete and covered then properties can be treated as concepts, as assumed in [15].

definition provides for a way of recording the relative importance of domains to symbols. Symbols are dependably associated with internal states of the conceptual system, so that a symbol can be viewed as labelling a concept, whether abstract or naming something in the environment. The importance function w summarises the associations which the individual makes between concepts and properties, and plays an instrumental role in the dynamics of conceptual spaces in Section 5.

Definition 4.2.2. *Importance* is a function $w: \{1, 2, \dots, l\} \times \{1, 2, \dots, Dom\} \rightarrow [0, 1]$ where $w(k, j)$ is said to be the importance of the domain Δ_j to the concept c named by the k th symbol. If the k th symbol names the concept $c(P_1, P_2, \dots, P_n)$ where P_i is in domain $\Delta_{j(i)}$, then $w(k, j(i))$ is also said to be the importance or salience of the property P_i to the concept.

4.3. Distances and similarity

Notions of similarity and distance have to be developed for objects, properties and concepts, and must take into account context. People are notoriously willing to judge the similarity of a pair of objects or concepts, but they are easily induced to vary this assessment between experiments. This is because similarity, and hence distance, depends on the context in which the assessment is made. The importance of context is well recognised in research concerning representations, with the favoured modelling mechanism being selective weighting of factors used to assess similarity. Context can be accounted in the choice of domains, ie. with a binary weighting. The following definition uses the Hausdorff distance introduced in Section 3 to define distances between concepts and between properties, since these are regions rather than points.

Definition 4.3.1 (*Contexts, distances between objects*).

- (a) A *context* is a set of domains. An object a is *defined in a context* if $\pi_i(a) \neq *$ for each domain Δ_i in the context. A concept $c(P_1, \dots, P_n)$ is defined in a context if the context includes the domains of each of the properties P_i .
- (b) The *distance* $d_C(a, b)$ between two simple objects a and b which are defined in a context $C = \{\Delta_{j(i)}, i = 1, \dots, K\}$ is $\sum_{i \leq K} d_{j(i)}(\pi_{j(i)}(a), \pi_{j(i)}(b))$.

The *distance* $d_C^H(A, B)$ between two regions A and B in the context C is the Hausdorff distance calculated using the distance d_C . That is,

$$d_C^H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \{d_C(a, b)\}, \sup_{b \in B} \inf_{a \in A} \{d_C(a, b)\} \right\}.$$

In particular, the distance between two properties or concepts can be defined using this function.

Defining the distance between simple objects or concepts independent of context, or when one or both objects/concepts are not defined in the nominated context, involves dealing with values $*$, denoting “not applicable” data. Thus, the clinician may try to compare a patient’s symptoms and signs with those of a case study anaemic they learned at medical school. But their patient does not have fatigue and is complaining about

backache which was not in the stereotypical profile. How should mismatch of conditions be dealt with? One approach is to simply ignore unmatched domains—though to ignore the backache could be irresponsible of the clinician. An even more drastic approach is to set the distance to infinity ie. treat the objects as incomparable. However, throwing out the possibility of anaemia because of reported backache is not a recommended diagnostic approach. More flexibly, a finite penalty on unmatched domains might apply, as in [1]. For example, cost might be proportional to the number of domains on which one object, but not both, is defined. The importance function w could be employed to support a more realistic costing.

Distance between complex objects is complicated not only by the possibility of values $*$, but also by the fact that the most appropriate match may involve a permutation of the levels (that is, the versions) of the base conceptual space. For example, object O may consist of a *yellow Volkswagen* in level 1, pulling a *boat* represented in level 2 on a *trailer* in level 3. Object O' may consist of the *boat* in level 1, and the *trailer* in level 2. Defining the distance between two such objects O and O' involves finding the best fit between the subobjects on which they are both defined, taking into account the cost of leaving parts of the objects unmatched.¹¹

It is evident that even when distances in domains are well understood, it may be difficult to relate them to the judgements that a conceptual system makes about distance between concepts in a given context.

As we have said, the natural definition of similarity of objects in terms of distances in conceptual space is one of the main advantages of the conceptual space representation. Like distance, similarity is a real valued non-negative bivariate function, call it s . The relationship between similarity and distance is largely one of definition, and the terms are used informally as reciprocal notions in many disciplinary areas. A commonly-assumed relationship is $s(a, b) = (1 + d(a, b))^{-1}$. While the nature of similarity is still debated in the cognitive literature (e.g., [17]), similarity is usually represented there as a negative exponential function of distance, and usually with an exponent of 1 or 2 (e.g., [38]). This functional form has been derived from experimental data in which distances d are relative scientific magnitudes rather than phenomenal distances. But there is still much to be done to understand the relationship between similarity and conceptual distance. For example, it may be desirable to judge things as having zero similarity even when they are only separated by a finite distance, for example, in assessing finger nail distortion against an exemplar. Nevertheless, reasonable postulates relating distance and similarity are:

- (1) similarity should be a maximum (at 1) when distance is zero,
- (2) similarity should be monotonically decreasing with distance,
- (3) similarity with $*$ should be zero for all points other than $*$.

The similarity of objects a and b might therefore be defined in the context C to be $\exp(-u \cdot d_C(a, b))$ where u is a sensitivity parameter. The distance between two objects would then be a multiple of the negative log of their similarity. The similarity of two concepts c_1 and c_2 defined in the context C would be $\exp(-u \cdot d_C^H(c_1, c_2))$.

¹¹ There will be unmatched parts whenever one of the objects has more components (occupies more levels) than the other, or when the objects take values $*$ in different domains.

4.4. An individual's conceptual space, associations, prototypes and categories

Definition 4.1.1 of base conceptual space as a product of a symbol space and a concept space associates each concept state with *every possible symbol*. Specifying the conceptual space to be the whole of $D_L \times D$ is analogous to specifying the language of a logic, or the structure of a neural network without the weights—it provides an infrastructure only. This infrastructure has to be refined through more precise linkage of non-symbol elements to symbols, just as a theory of logic might capture an individual's beliefs

Ideally, each point x in D (other than $*$) would be associated to just one point in D_L , namely the vector which specifies the applicability, in the range 0 to 1, of each of the symbol dimensions to x , or equivalently, specifies how representative the concept x is to the symbol vector. For example, “palpitations”, “pallor” and “spoon shaped nails” might be assessed as being very typical of a patient with iron deficiency anaemia but less typical of leukaemia, and this would be reflected in the associations between the points associated with these qualities in D and the symbol dimensions for “iron deficiency anaemia” and “leukaemia”. At worst, a point in D might be associated with a *range* within $[0, 1]$ for each symbol dimension, to allow for imprecision in the notion of applicability. Most symbol dimensions will be inapplicable to any object and so take the value zero (that is, the $*$ value on these dimensions).

We define an *individual's conceptual space* $A(C)$ to be the subspace of $D_L \times D$ which only contains pairs (s, x) of symbol dimension vectors s and points x in D in which s reflects beliefs about the appropriateness of the symbols to the points. This subspace describes how each symbol is associated with the region (including singletons) to which it refers. The associations will vary according to the individual, and will vary as the individual learns. For communication, as we have said, associations that different individuals make must have enough commonality to support some shared understanding of the referents of the symbols exchanged. Thus the form of the association cannot change too fast. As learning is not the focus of this paper, we assume that the associations are fixed (at least, on the time scale of the dynamics considered later).

Properties can now be described in terms of the individual's conceptual space. Suppose a symbol indexed by $k \in \{1, 2, \dots, l\}$ names a concept $c(P_{k,r(1)}, \dots, P_{k,r(q)})$, where $P_{k,r(j)}$ is a property in domain $\Delta_{r(j)}$. Then $P_{k,r(j)}$ is the region in this domain associated with this named thing by the individual.¹² Thus, it is reasonable to propose that $P_{kr(j)} = \{\pi_{r(j)}(y) : y \in A(C); \pi_k(y) = 1\}$. For example, suppose that the k th symbol dimension represents the concept “anaemic patient”. Then associated properties $P_{k,j}$ might be the region in the colour domain, within the subdomain of complexion colours, considered pale; or the region in the fatigue domain which represents lassitude. These properties might be the projection of all the points in the individual's conceptual space in which the label “anaemic patient” was fully applicable, and so would include the projections of any remembered cases of patients classified as anaemic. Of course, a patient can be anaemic without being pale, or indeed without exhibiting any symptoms or signs, if they have decreased red cell volume but have physiologically compensated for it. However,

¹² Here and elsewhere, if $P_{k,j}$ is defined, it is assumed to be non-empty.

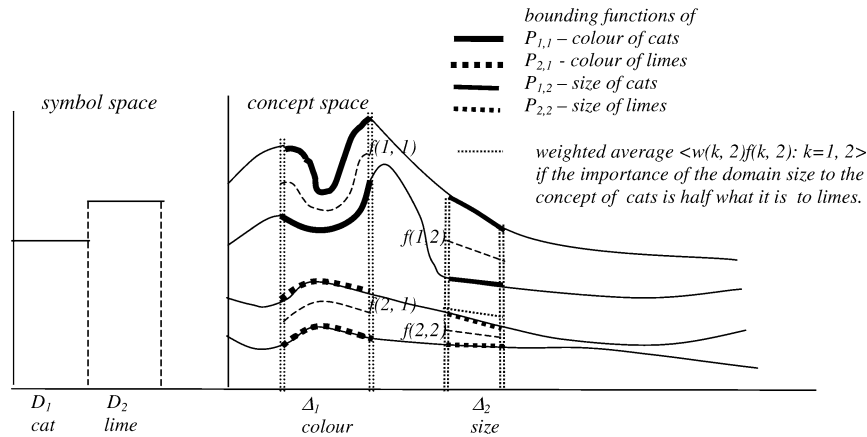


Fig. 5. Notation for prototypes. $f(k, j)$ denotes the prototypical value in the domain Δ_j for the concept labeled with the k th symbol, viz. the centroid of the region $P_{k,j}$ in the domain associated by the individual with the concept. In this example, points in conceptual space are the family of functions described in Figure 4. Any two functions define a region, namely the set of functions which lie between them, where $B(f, g, h)$ if and only for all x either (a) $f(x) < g(x) < h(x)$ or $f(x) = * = g(x) = h(x)$ or (b) $f(x) > g(x) > h(x)$ or $f(x) = * = g(x) = h(x)$.

normal conditions will not be explicitly associated with anaemia in any reasonable conceptualisation.

Define the *prototypical value* $f(k, j)$ in the domain Δ_j of the concept labeled with the k th symbol dimension to be the centroid¹³ of $P_{k,j}$, computed as in 3.1.8(b). A characteristic such as convexity is needed to ensure that the centroid is actually in the region, although it could plausibly be argued there is no reason that the prototype should be a *member* of the region, only that it be the best *representative* of the members of the region. To illustrate the notion, Fig. 5 depicts prototypes in the example setting used in Fig. 4, in which points in conceptual space are piecewise continuous functions.

Clearly the same prototypical value $f(k, j)$ can represent many different regions. In the limit, it represents the point set which is its own value, which could occur when a concept has been learned through explicit tuition, rather than having been experienced. Thus to a novice an anaemic patient may be associated with a prototypical value for tongue texture corresponding to a medical web site picture. The use of prototypes is critical to the manipulation of conceptual spaces, as will be evident when we deal with dynamics in Section 5.

Given a set of symbol indices K , the weighted average of the associated prototypical property values $f(k, j)$ for $k \in K$ can in turn be defined as a centroid using 3.1.8(a). If the weight on $f(k, j)$ is $w(k, j)$, denote such a centroid by $\langle w(k, j) f(k, j) : k \in K \rangle$.

While prototypes can be defined as centroids of regions, they can also define regions, because of the tessellation Theorem 3.1.7 which partitioned a space into n disjoint convex

¹³ This assumes $P_{k,j}$ is compact; if not, the prototype can be defined as the centroid of a compact subset of $P_{k,j}$, or of the compact closure if this exists.

spaces around a set of prototypes p_1, p_2, \dots, p . Thus a set of patients can be divided into normal and anaemic, and the latter subdivided into those with iron-deficiency anaemia and those with chronic disease anaemia, and so on. Case studies might provide the prototypical examples of the diseases.

Alternatively, the prototypes used in defining properties $P_{k,j}$ as convex regions might be the centroid of a finite set of *exemplars*, that is, as the prototype of a *category* whose members instantiate the concept labelled by k . A category is a collection of objects or things sharing some conceptual likeness. Gärdenfors defines a category to be a set of objects instantiating a defining set of properties $\{P_1, P_2, \dots, P_k\}$; however, categories may be formed on the basis of theories, similarity to exemplars, or distance from prototypes.¹⁴ We suggest that, given disjoint categories of objects classified by theories, perhaps, or via scientific measurements, then exemplars of each category might be used to form prototypes, computed as the centroids 3.1.8(a). (Such a process might involve scientific studies establishing the symptoms and signs of a disease, and exposure to cases teaching the clinician to recognise these symptoms and signs as indicators of the disease.) The prototypes of categories of exemplars in turn could be used to determine properties of the category, using Theorem 3.1.7. While it is desirable that the prototype of a property coincide with the centroid of the region defining the property, even in Euclidean domains tessellation creates regions in which the centroid computed using Voronoi tessellation can be distant from the prototype calculated from the exemplars. So forming a property $P_{k,j}$ in this way means $f(k, j)$ may not be the centroid of the exemplars of the concept labelled by k .

5. Dynamics of conceptual spaces

5.1. Introduction

This section applies conceptual spaces to two key aspects of problem solving, namely differentiation and composition, and shows how structures can be derived through dynamics.

We have said that representations may be thought of as states which support a system's purposeful interaction with its environment. Conceptual spaces must be equipped with an algorithm which allows the state of the system to change in response to input or to internal computation, in analogy with connectionist and symbolic systems. Each point in a conceptual space is a potential conceptual state: being "what you're thinking about" if the point is inside active areas of attention, and being "what you know" if it is outside active areas of attention. Problem solving is a progression of thoughts which are, sometimes at least, related in the systematic fashion that we call reasoning. To support problem solving, a mechanism is needed to move from one conceptual state to another in response to external information or internal reasoning strategies. Problem solving highlights the need for multiple levels of representation to depict complex objects, because several conceptually-distinct objects may need to be manipulated at the same time.

¹⁴ The role of prototypes, or even similarity to exemplars, in categorisation remain disputed [36]. There are many reasons for defining a category as a problem solving or information-ordering device, and thus many mechanisms are used to categorise things (see, for example, [25]).

Introducing dynamics into the formalism requires a time parameter. At any time t , an individual's conceptual state is described by a point in conceptual space. The dynamics describe the movement of this function in conceptual space over time. We adopt a discrete time perspective, as one physiologically as well as computationally motivated by system reaction times. The time interval between temporally adjacent states is taken to be one unit.

The dynamics can be phrased generally as state transitions or as rules. In either case, state changes are effected through the associations between substates in the symbol and concept spaces. The transition rules or difference equations which govern state changes are phrased in terms of regions and distances, and so are embedded in the geometry of conceptual spaces. The active areas in each subsystem are determined by a mask which leaves open what we call the *area of attention*.

Management of attention is crucial to problem solving. Two assumptions about human attention are “nearly axiomatic” in cognitive psychology, according to Barsalou [3], namely, information is isolated by being focused upon in perception, then, to a “very high likelihood”, it is stored in long-term memory. He says “*From decades of work on attention, we know that people have a sophisticated and flexible ability to focus attention on features . . . , as well as on the relations between features*”. The mechanisms by which attention is controlled are not clear, although areas of attention change rapidly. Control of attention is also an important research area in robotics and computer vision. In ART-like networks, attention is modelled using a gain control which partially controls interactions between the upper (symbolic) and the lower concept (pattern) layers (e.g., [5–7,29]).

Our system uses an attention buffer, that is, a sequence of attention areas which are stepped through as the state in each area of attention stabilises. The buffer is filled by an algorithm which takes as input the current state of the symbol subspace. Multiple symbols active on the one level result in *composition* of concepts, and symbol activity at multiple levels results in the *selection* of one of the levels. These two modes of operation, combine and select, are evident in a wide range of processing systems. Neural networks, whether artificial or biological, exhibit inhibitory or reinforcing modes of activation dependent on whether concepts combine or compete. Paredis [31] has argued that problem solving is the interaction of selection and combination processes at a micro level. In his work, these processes originate from the application of simple rules to pairs of elementary solution constituents. In yet another setting, Humphreys et al. [22] define “Choose” and “Intersection” primitives for human memory access systems to perform these functions.

The symbol space exchanges information with the external world, and plays a key role in initiating sequences of state changes. It is not, however, necessary to suppose some higher level authority is in control or interpreting the results of state changes. Rather, concept states may change in response to the symbol state, and symbol states may change in response to the concept state, in an endless, aimless processing loop.

In this section, the form of state transitions is explored and the algorithm for setting attention is presented, along with control principles. The formation of composite concepts using the dynamics is then described. The final section illustrates the dynamics on a typical example of categorisation, which, as we have said, is the key problem-solving activity at the conceptual level. The example draws out the inherent relationship between geometry in the space and the trajectory of conceptual states during problem solving. Section 6 presents specific examples of conceptual spaces and their dynamics.

The following terminology and notation will be used:

Definition 5.1.1.

- (a) An individual's conceptual state at time t is denoted $c(t)$. That is, $c(t) \in A(C)$.
- (b) If k indexes a symbol dimension, $c(k, t)$ denotes the projection onto the k th dimension $[0, 1]$ of D_L of $c(t)$ and is called the *activity level* of the k th symbol dimension. We also say that symbol k is *active* at a point $x \in C$ if $c(k, t)$ is greater than some minimum value.¹⁵
- (c) $c_i(t)$ denotes the projection of $c(t)$ onto the i th concept domain, viz $c_i(t) = \pi_i(c(t))$. The notation differs from that for symbol space to make it obvious whether projections are onto symbol or concept spaces.
- (d) Given an element x in a concept domain Δ_s , let $\pi_s^{-1}(x)$ denote $\{y: \pi_s(y) = x, \pi_r(y) = * \text{ for } r \neq s\}$. This can be defined, because concept domains are separable and C is the product of concept and symbol space. We refer to $\pi_s^{-1}(-)$ as “the state over domain s ”.

5.2. Dynamics

Dynamics in conceptual spaces are based on five key notions:

- (a) The current conceptual state is represented as a point in conceptual space C^v which follows a non-deterministic trajectory as a result of external input (see (d)), and internally-derived activity (see (c)).
- (b) For most problems, only a relatively few concepts are relevant, and problem solving needs to focus on these. A time-varying buffer is used to carry a prioritised list of windows or areas of attention, which are regions in base conceptual space. The head of the buffer specifies the current area of attention, which will not necessarily change between time periods.
- (c) Changes in the current state in conceptual space occur only in attention areas. Changes can be seen as transitions of a recurrent network or dynamical system controlled by the attention buffer. Alternatively, they can be seen to be consequences of rules activated when their preconditions are a “close enough” match to the current area of attention and/or the current state in the current area of attention. (This is analogous to a database update in which the area to be updated is identified through the buffer.)
- (d) The conceptual state receives external inputs primarily through the symbol dimensions, and produces output through these dimensions. This can be viewed as part of a two-way symbolic communication, possibly with “higher” systems although also possibly with peer level systems. There can also be communication with a subconceptual systems which might alter the concept state. Recall Fig. 1.
- (e) When an attention buffer is emptied while a symbol or set of symbols L is still active, an operation $SetAttention(L)$ sets the initial state and refills the buffer with the areas of attention which are most salient in L (in a sense discussed below). If no

¹⁵ The minimum value may be globally set. It may be zero.

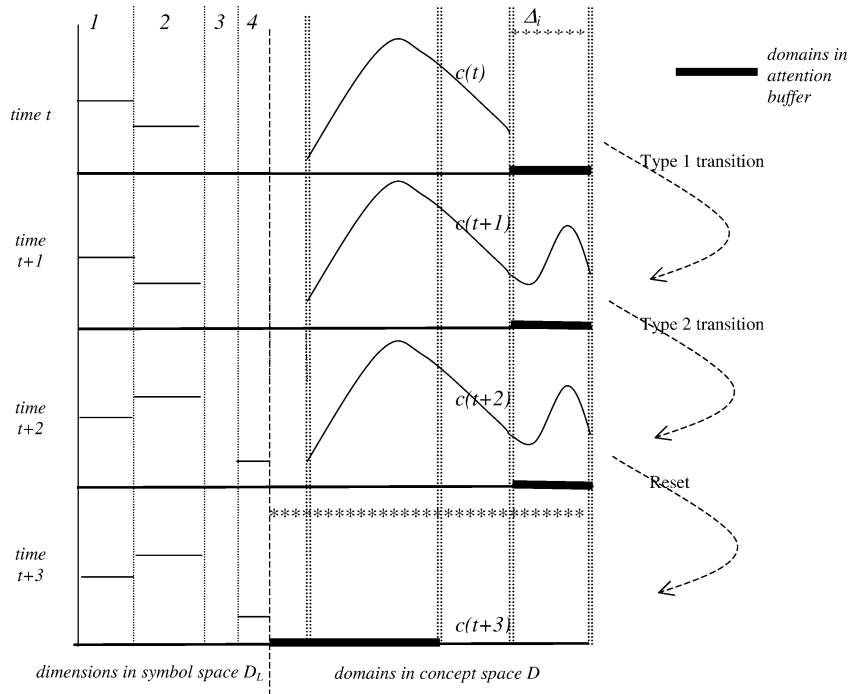


Fig. 6. Dynamics in conceptual space. Four epochs are shown. Regions of change are dictated by an attention buffer, depicted schematically here as a dark bar indicating the relevant domains in D . Transitions of type 1 set states in areas which had projected onto *, as in the transition $c(t) \rightarrow c(t+1)$. Transitions of type 2 change values in symbol dimensions, as in the transition $c(t+1) \rightarrow c(t+2)$. Reset refills the attention buffer and sets current state to have existing values in symbol space, and undefined values in concept space, e.g., $c(t+2) \rightarrow c(t+3)$.

symbols are active, a repair or training situation is entered. The management of the area of attention through this operation is central to problem solving.

There are two fundamental types of transitions—symbol activity triggering concept state transitions, and vice versa—and control devices including the setting of the attention buffer. These are described below. These operations are sufficient to perform categorisation, as is shown in the next section. Fig. 6 illustrates typical results of actions of the operations, using the example of a conceptual space as a space of piecewise continuous functions. An example of how forms of transitions can equip the space with particular problem solving strategies is provided in Section 6.

Two important points: Firstly, the attention buffer is always phrased in terms of the base conceptual space, and the region specified in $Att(0)$ is the region of attention for *all* levels. Transitions occur within a level, so unhelpful complexity is avoided in the following exposition by treating $c(t)$ to be scalar rather than vector.

Secondly, changes to the current state $c(t) \rightarrow c(t+1)$ are specified through changes to values in domains, and domain values do not change unless specified. This is possible because of our assumption that conceptual space covers the product of domains and symbol dimensions. Except in the special case in which the concept space D is also complete

(i.e., is a product of the domains), specifying what happens on projections to domains does not fully determine the transition from state $c(t)$ to state $c(t + 1)$. That is, the transition in the period $[t, t + 1)$ may have a non-deterministic element within conceptual space; only the projections onto the domains are fully determined.

5.2.1. Transitions

The first fundamental type of state transition, call it type 1, changes the state in concept space, i.e., changes some values $c_i(t)$. Type 1 transitions set the current state in the area of attention, *provided that the current state is not already defined there* (that is, provided $c_i(t) = *$) and provided there are active symbols associated with this area. A type 1 transition therefore *initialises* part of the concept state space. The second transition type changes the state in symbol space, i.e., changes some values $c(k, t)$. Type 2 transitions feed back activity to the symbol dimensions associated with active regions in conceptual space *D if these regions are also in the area of attention*.

Rich use of geometry may be involved in defining the scope of transitions or their effect. For example, a type 1 transition might depend not only on which symbol dimensions are currently active, but also on whether the prototypical value (centroid) $f(k, i)$ for the concept associated with the k th symbol on domain Δ_i is in the area of attention. Then the state change may depend on another calculation of a centroid, using the combined weights of the importance w of the domain to the concept, and the activity of the symbol labelling the concept. A type 1 transition might therefore have the following equation, which initialises the state on a domain Δ_i to be the centroid of the set of prototypical values (in this domain) of active symbols, provided these values are in the region of attention:

$$c_i(t + 1) = \langle c(k, t)w(k, i) f(k, i) : f(k, i) \in \pi_i(Att(0)) \rangle \quad \text{if } c_i(t) = *. \quad (5.1)$$

In the situation depicted in Fig. 6, the value of $c(t + 1)$ in domain Δ_i would thus be the weighted centroid of the prototypical values in this domain of the active symbols in symbol dimensions 1 and 2. If Δ_i were fatigue and the active symbols were “Patient X” and “iron deficiency anaemia”, a representation of fatigue is activated which is a compromise between that of the patient and the typical fatigue level of those with the disease, provided both the prototypical values for fatigue of Patient X and fatigue associated with iron deficiency anaemia are in the area of attention.

Type 2 transitions raise the activity level of a symbol which has a prototypical value near the current state, provided this part of the current state is in the region of attention. The adjustment to the activity level, $c(k, t + 1) - c(k, t)$, would again in general be a function of the importance of the property to the thing labelled by the k th. symbol. For example, a fatigue level very representative of tiredness of iron deficiency anaemics might still only weakly increase activity in the anaemics’ symbol dimension because fatigue is not an important distinguishing domain here. An example of such a type 2 transition is:

$$c(k, t + 1) = c(k, t) + G(w) \quad \text{for each } k \in \{j : d_i(c_i(t), f(j, i)) < \delta \ \& \ c_i(t) \in \pi_i(Att(0))\}, \quad (5.2)$$

where $G(w)$ is some geometrically-based real valued function of the importance $w(k, i)$ of the i th dimension to the k th symbol. G might be related to the distance between the projection of the current state on a domain and prototypical values for the k th symbol,

perhaps modulated as a similarity judgement to bring it into the range [0, 1]. Specifically, $c(k, t + 1) - c(k, t)$ might be $w(k, i) \exp(-d_i(c_i(t), f(k, i)))$, or more generally, for some positive constant $\lambda \leq 1$,

$$c(k, t + 1) = c(k, t) + \lambda \sum_{ci(t) \in \pi_i(Att(0))} w(k, i) \exp(-d_i(c_i(t), f(k, i))). \quad (5.3)$$

Of course, $c(k, t + 1)$ cannot exceed 1, so fully extended symbols are not affected by such transitions. If (5.3) holds, the type 2 transition that Fig. 6 depicts between time $t + 1$ to $t + 2$ could only occur if the concept labelled with the 4th symbol had a prototypical value in domain Δ_i close to $c(t + 1)$ —for example, if this concept were leukaemia and fatigue symptoms for leukaemia were similar to those of Patient X and of sufferers of iron deficiency anaemia. While the precise form of the transitions will depend on the application, type 1 and 2 transitions would be expected to be a function of the importance $w(k, i)$, as was the case in (5.1) and (5.3).

5.2.2. Control and SetAttention

There are two general control principles for conceptual spaces. The first control principle P_1^C says move attention ($Att(0)$) to the next element ($Att(1)$) when activity dies out in the current region. The second principle P_2^C says fill the buffer when it is emptied, and do so according to the current state of symbol space. Specifically:

P_1^C : If the state does not change in a unit time period then $Att(i) \leftarrow Att(i + 1)$, $0 \leq i < s$.

P_2^C : If Att is empty then $SetAttention(L; \eta)$ where L is the complex concept formed from the dimensions k in symbol space which have activity level $c(k, t)$ greater than an externally set level η (or are the dimensions remaining after a cut-off calculated to limit the number of symbols).

SetAttention sets attention areas and the initial state for the next round of processing. The attention areas will apply across all levels, that is, across all the copies of the base conceptual space. However, the operation of *SetAttention* itself takes account of the levels of its input. Recall that levels are used to represent conceptual subcomponents, such as a “flat tyre” on a “car”, or “fingernails” of “Patient X”. The algorithm assumes that when the structure of complex objects is presented through active symbols set at different levels, then the objective will be to distinguish them. So it tries to find properties on the domains which best discriminate concepts associated with the symbols. Multiple active symbol dimensions at the one level portray, in contrast, a conceptually indivisible thing such as a “anaemic patient” or “flat tyre”. The algorithm therefore assumes that the objective will be to find the areas of commonality of the symbols, to isolate the thing to which the multiple symbols refer. So the algorithm tries to find properties which the active concepts or objects have in common. These properties are then fed into the multi-layer comparison. All cases require a measure of the distance between properties, which we implement using the Hausdorff distance d^H defined on the relevant domains.

SetAttention is presented in Fig. 7, for the cases (a) all active symbols are on the same (base) level, and (b) each active symbol is on a different level. If there are active symbols at more than one level, and multiple symbols such as “pale” “anaemic patient” are active in at least one level, then *SetAttention* first forms the composite concepts at each level, as in

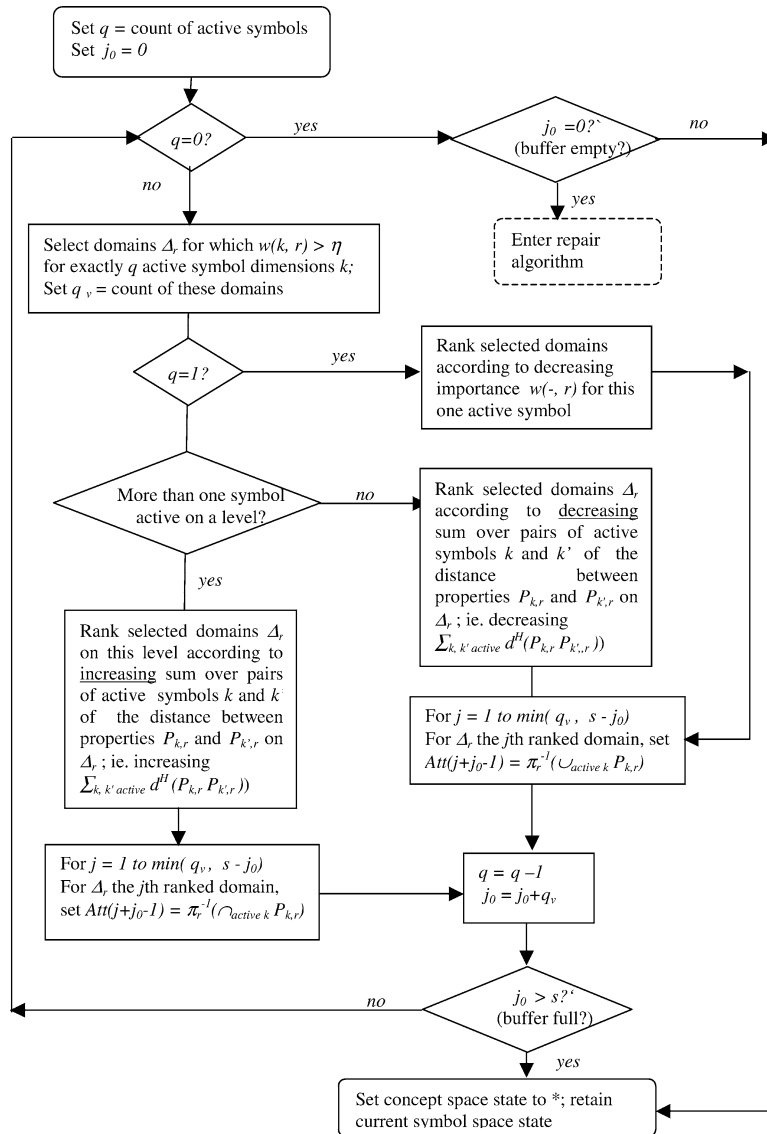


Fig. 7. *SetAttention* algorithm. Input is the active symbol dimensions and a cutoff η . For simplicity it is assumed either each active symbol is at a different level (so goal is to differentiate) or all are at the same level (goal is to find commonalities). *SetAttention* finds properties or domains which are associated through the importance function w with as many of the input symbols as possible. It then orders properties by how common they are to the labeled concepts on the same level, or it orders domains by how well they discriminate labeled concepts on different levels, in the sense of separating the regions on the domain associated with each concept. If the buffer is not full after one such attempt, the algorithm tries again on properties or domains which are not associated with as many of the input concepts. If it fails to find any associations between any of the input concepts at this level, it will enter a repair algorithm. If only one symbol dimension is active, *SetAttention* finds the most important properties for the concept using the importance function w .

the algorithm for case (a). Instead of putting the ranked regions, which are the non-trivial intersections of properties of the active symbols, directly into the attention buffer, it uses them for the next stage of processing. This is essentially algorithm (b), with these regions used in place of the regions (properties) presumed to have been associated with the one symbol active at each level.

5.2.3. Overview of the formation of composite concepts

After *SetAttention* completes, only symbol dimensions are active. Suppose these symbols refer to a composite concept, such as “worried person” “leukaemia sufferer”, and so are at the same (base) level. A type 1 transition then sets the current state in the initial area of attention, which is the region in the concept domain which has *most in common* for the concepts named by the active symbols. The value taken will be the centroid of the prototypical values of these concepts on that domain, according to (5.1). This change allows type 2 transitions to raise activities of symbols labelling concepts which have prototypical values near the centroid, using the definition of distance in that domain. If the property considered were widely-used, such as “large” on a size domain, or “frowning” on a facial expression domain, there may be many irrelevant concepts activated. For example, if *SetAttention* had operated on the symbols “large size” and “animal”, at this stage symbols of large cities, large cars etc. may all be activated along with examples of large animals. If the facial expression property “frowning” was shared by “worried person” and “leukaemia sufferer”, then frowning people of various ilk might be recalled and their name symbols activated, through (5.3). These, however, will not cause further type 1 transitions because attention is on the initial domain (a frowning facial expression, say).

Therefore, under the general control principle P_1^c , attention moves to the next region *SetAttention* stored in the attention buffer. This refocus allows another type 1 transition, setting the state over the new domain to the centroid. This time, the centroid may not only be calculated using the prototypical values of the concepts associated with the original symbols (“worried person” “leukaemia sufferer”, say), but also prototypical values for other active symbols, provided they are in the region of attention. A type 2 transition will then again raise activity levels of any symbol associated with a concept “near” to this prototype, using (5.3). Symbols connected with worried leukaemia sufferers—such as the names of patients—are likely to be at highest activation levels, having been raised by each of the type 2 transitions. Other symbols irrelevant to the composite concept may be active, but their activity levels would, in general, be raised by fewer of the type 2 transitions, and so would be likely to be lower. The cycle of type 1/type 2 transitions is repeated till the buffer empties.

SetAttention is called again, under the general control principle P_2^c . The symbols above the cutoff threshold, for example those naming the worried leukaemia patients, are activated on the same level, and the properties that they share are ranked by *SetAttention*. It is at this stage that a conceptualisation of the indivisible concept described by the multiple symbols, e.g., “worried leukaemia sufferer”, becomes available for further processing. The regions (properties) derived by *SetAttention* can be captured as a new concept, and their ranking used to define the importance function w on the composite concept.

5.3. Categorisation example

This section shows how the *SetAttention* operation, the control principles and the two transition types are sufficient to solve general categorisation problems. Descriptions of the objects to be categorised can be presented through the symbol space D_L (mimicking input through the symbolic system), or through the concept space D (mimicking perceptual input, for example visual input through the pre-conceptual level). We only consider the former case here.

Typical categorisation problems involve a question such as *What type of anaemic patient is Patient X: leukemic or iron deficient?*, or *Does Patient Y have leukaemia or an iron deficiency anaemia?*, where *Patient X* and *Patient Y* are known to the person being asked the question, or rather, the conceptual system has some associations between the symbols and conceptual states. Alternatively, the question might be *Does this suggest leukaemia or an iron deficiency anaemia?*, followed by presentation of a list of signs, symptoms and test results for *Patient X* or *Patient Y*. Solving such problems correctly requires “enough” overlap between (i) the properties describing *iron deficiency anaemia* and *leukaemia* in the conceptual space of the problem solver and (ii) their understanding of the properties describing the symbols *Patient X* and *Patient Y* in the first formulation of the problem—in a sense that will become clear as we work through the example. In the second formulation, the overlap must be between (i) the properties describing *iron deficiency anaemia* and *leukaemia* and (ii) the stated or observed properties of the anaemic patients.

In the following, suppose the external input which provokes conceptual activity is as shown in quotations. Suppose further that possible categories into which objects are to be placed are k_1, \dots, k_g . Finally, suppose transitions are described by Eqs. (5.1) and (5.3), where the importance function w is known to the system. Then categorisation proceeds with these steps:

“*Is the following an example of k_1 or \dots or k_g* ” (e.g., “*Is the following a case of leukaemia or iron deficiency anaemia?*”).

- (a) The symbol dimensions k_i are activated from outside the system at multiple levels, because this is a comparison operation.
 - (b) The control rule P_2^C invokes *SetAttention*, which fills the buffer with areas in domains that best discriminate the categories k_1, k_2, \dots, k_g . Each of the symbols is active at a different level, and $c(t)$ is set to * everywhere else.
- (i) “*Object O*” (e.g., “*Patient X*”).
- (c) The input from outside the system sets the current state so that just the symbol dimension corresponding to the object O is active, call it o .
 - (d) If the conceptualisation of O projects non-trivially onto a property $P_{o,i}$ in the i th domain which has centroid $f(o, i)$ in $Att(0)$ then the current state over domain i is set by (5.1) to $f(o, i)$ at each of the levels. Else the next buffer item is selected.
 - (e) If a type 1 transition has been made, it is because the prototype of property $P_{o,i}$ is in an attention area. This is a region on the domain which best discriminates the classes, according to *SetAttention* (Fig. 7). Given the form of (5.1), the change to the state is likely to have brought it close to the prototypical values on this domain of one of more of the classes k_i . If so, the change in the state will allow a type 2

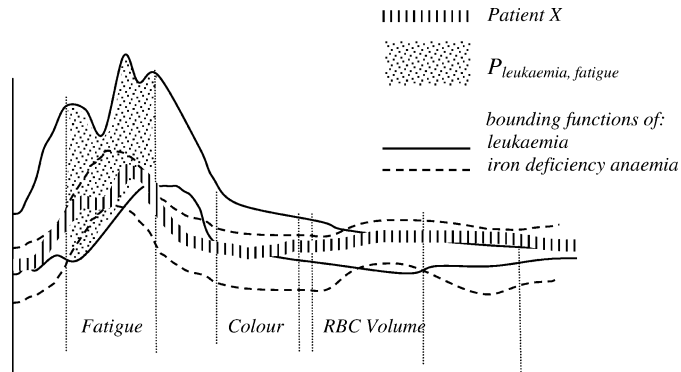


Fig. 8. Role of metric concepts in categorisation. Conceptual space is represented as a family of functions, as in Fig. 4. Suppose D has domains *Colour*, *Fatigue* and *Red Blood Cell Volume*. The figure depicts disease concepts, labelled *leukaemia* and *iron deficiency anaemia*, with overlapping properties (functions between a pair of bounding functions) on these domains. The conceptualisation of the symbol *Patient X* is also a region (i.e., a set of functions) which lies wholly within the concept of *iron deficiency anaemia* but intersects with *leukaemia* on all 3 domains. The categorisation of *Patient X* will depend on where the centroids of the properties associated with *Patient X* lie in relation to properties such as $P_{leukaemia, fatigue}$ associated with the diseases. It will also depend on distances between centroids.

transition, Eq. (5.3), raising activity $c(k, t + 1)$ for one or more of the class symbols $k \in \{k_1, \dots, k_g\}$.

- (f) Steps (d) and (e) are repeated until the buffer is emptied. If there have been no state changes to conceptual space, the system moves to a repair/training situation through the control rule P_2^c . Otherwise, one or more of the class symbols will be active (and possibly other symbols as well).

Note that the “Patient X” symbol will still be active at the end of processing, along with, say, an iron deficiency anaemia dimension symbol, and a lower activation on the leukaemia symbol, if conceptual space is as depicted in Fig. 8. These activation levels provide the categorisation mechanism. The active symbols are used by *SetAttention* in the next round of processing, unless there is some external input: specifically, the algorithm will try to set the attention buffer to common properties of the active symbols.

OR (ii) “an individual with properties O_1, O_2, \dots ” (e.g., “the anaemic patient is pale, suffering headaches, has fingernails which are concave, etc.”).

- (c) A symbol corresponding to each of the properties is externally activated, all on the same level since they are part of a composite object. For simplicity assume the property Q_i associated with the symbol O_i labelling that property is on domain i .
- (d) If the prototype for a property Q_i is within an entry in the attention buffer it must, because of the way *SetAttention* has selected areas, overlap one or more of the properties used to discriminate the classes. If so, when the attention buffer head reaches this entry, the type 1 transition (5.1) sets this area in the current state to be a prototype of Q_i .
- (e) As in (i)(e) above, a type 1 transition (5.1) occurs in (d) because the changed state is likely to be near a prototypical value for one or more of the properties of the classes.

So it is likely to allow a type 2 transition (5.3) to feed back activity to one or more of the category symbols k_i which have similar prototypical values to the properties of the individual.

- (f) Steps (d) and (e) are repeated till eventually the state c attains prototypical values for each of the properties of the object which have “enough” overlap with the discriminating properties to trigger their associated transitions, and classes with more similar prototypical values on the domains have higher activation levels.
- (g) Thus a ranking of the class symbols is possible through their activity levels.

In this case, the property symbols O_i and the iron deficiency anaemia/leukaemia dimension symbol(s) are, in the absence of external intervention, left active when the buffer empties, for use by *SetAttention* in the next round of processing.

There are three important points to note about the example. Firstly, there is no need for high level interpretation of the solution on conclusion of the categorisation. The current state of the symbol space may simply be passed to *SetAttention* in an endless processing loop, into which may be injected external changes of state, and interrupts to current processing causing resetting of the attention buffer using the current symbol space. Secondly, the procedure is an example of an anytime algorithm, in the sense that it can be interrupted at any time after a problem has been presented, and a tentative solution is available for further processing in the current state of symbol space. In particular, this means that the size of the attention buffer (the number of properties which are being used to categorise each object) may not be as critical as the time given to the task before interruption. Thirdly, the categorisation depends intimately on the geometry of conceptual spaces, as this is used in the setting of the attention areas and in the determination of what transitions are made, as well as in the associations between symbols and concept state.

6. Examples of conceptual spaces

This section illustrates the definitions of conceptual spaces and the dynamics described in the previous section. First, dynamical systems are represented as conceptual spaces, including emulation of their dynamics. This, however, provides limited demonstrations of the power of conceptual spaces, because there is almost no geometry to exploit. Our framework is then applied to the geometric representation of conceptual spaces developed by Gärdenfors [15], which has been the foundation for our extended definition of conceptual spaces. Inductive inference is the key reasoning activity flagged for the geometric representation, and this is implemented here.

Another formulation, a two-dimensional version of the example in Fig. 4 which we call “voltage maps” and which are reminiscent of topographical maps in cortex, is developed in [2].

6.1. General dynamical system implemented as a conceptual space

Dynamical systems are normally approached from a perspective of continuity, even when simulated on discrete computer systems. Their use in neurobiology and psychology

is linked with the activity of pre-conceptual level neural processes, rather than higher level processes, and it is most natural to model input to the conceptual space representation of a dynamical system as coming from the pre-conceptual level (i.e., directly to the concept space dimensions). To input through the symbolic level is cumbersome. The design choice made here is to model the dynamical system as a family of vector fields on a manifold X , then to use the activity level of a symbol to denote the real value in a component of \mathbb{R}^n at a point on the manifold. Thus, the number of symbols is n times the cardinality of X .

Definition 6.1.1 (*Dynamical system as conceptual space*). A formal definition of a dynamical system as a conceptual space that allows for input through the symbol space is:

- $D - \{*\}$ is a finite family of continuous vector fields \mathbb{R}^n on a manifold X , and the index set of concept space dimensions Γ is X .¹⁶
- For each dimension x , $D_x = \mathbb{R}^n \cup \{*\}$ with the usual Euclidean distance, and $\pi_x(y)$ is the value of the vector field y at x ; betweenness on D is defined by betweenness in \mathbb{R}^n at every point on X so that $B(a, b, c)$ if and only if $B(a(x), b(x), c(x))$ for each $x \in X$ and for all x, y in X , $B(a(x), b(x), c(x)) \Leftrightarrow B(a(y), b(y), c(y))$.
- There are no integral domains.
- Corresponding to each concept space dimension x there is a set of n symbol dimensions.
- $v = 1, s = 1$.

Specification of the dynamics is through linking symbols to states in D , and through difference equations $f(-, t + 1) - f(-, t) = \psi(-, t)$. The buffer is set to depth 1, and transition rules are used to switch attention amongst non-symbol dimensions.

Suppose X has discrete approximation $X' = \{x_1, x_2, \dots, x_N\}$ (see footnote 13) and dimension D_i is indexed in X by x_i . Let $k_{i,j}$ denote the symbol dimension corresponding to the j th component of the input field in the i th dimension. Let ξ be an isomorphism between the reals and the open interval $(0, 1)$. The weight function $w(k_{i,j}, x) = 1$ if and only if $x = x_i$ and $f(k_{i,j}, x_j) = \xi(c(k_{i,j}, t))$.

Then the type 1, type 2 and update transitions are:

$$\left\{ c_i(t + 1) = (\xi(c(k_{i,j}, t)))_j; Att(0) = D_i \text{ modulo } N+1 \right\} \\ \text{if } D_i = Att(0) \ \& \ c(k_{i,j}, t) > 0 \ \& \ c_i(t) = *, \tag{6.1}$$

$$\left\{ c(k_{i,j}, t + 1) = \xi^{-1}((c_i(t))_j), \ j = 1, \dots, n \right\} \text{ if } c_i(t) \in \pi_i Att(0), \tag{6.2}$$

$$\left\{ c_i(t + 1) = c_i(t) + \psi_i(t); Att(0) = D_i \text{ modulo } N+1 \right\} \\ \text{if } D_i = Att(0) \ \& \ c_i(t) \neq *. \tag{6.3}$$

The input to the system is a vector $\{ \{a_{i,j}\}_{j=1,n} \}_{i=1,N}$ entered as a positive activity level $\xi^{-1}(a_{i,j})$ on each of the symbols $k_{i,j}$. *SetAttention* tries to find common properties and

¹⁶ In a discrete representation the dimensions would be a finite subset $X' = \{x_1, x_2, \dots, x_N\}$ of X , and there would be a family of approximation functions Ψ defined on X' such that for any $\varepsilon > 0$ and any vector field v on X in the finite family, there is $v' \in \Psi$ and some extension v^* of v' to X satisfying $|v - v^*| < \varepsilon$ and $\sum_{i=1,N} |v(x_i), v^*(x_i)| < \varepsilon$. In the discrete representation, D is the family Ψ defined on X' .

will find that each $k_{i,j}$ will be associated with the whole of D_i . So it will end up simply putting D_1 into the attention buffer $Att(0)$. The type 1 transition sets the current state at x_1 to $\{a_{1,j}\}_{j=1,n}$ and changes the attention to D_2 . Another type 1 transition sets the current state at x_2 and so on until the state c is set with the correct input at each node. (This is the initial stage if the input is received directly at the dimension x_i .) Update transitions are then made sequentially, feeding back results to the symbolic level system through the type 2 transitions.

6.2. Geometric mode of conceptual space representation

Gärdenfors [15] emphasises the value of a geometric interpretation of perceptual concepts, because of the intrinsic notion of distance. This can be used, inter alia, in the important reasoning activity of induction, that is, of generalisation based on correlation between properties. The essential role of induction is to establish *connections* among concepts or properties *from different domains*. The dynamics of induction as generalisation involve the replacement of unknown values in dimensions with properties assigned on the basis of correlations.

The preferred structuring of the conceptual space is as a Cartesian product. There is strong evidence for non-separability of many perceptual dimensions such as those involved in colour: a group of integral dimensions is supposed to exist when a Minkowski metric of degree greater than 1 (typically, the Euclidean metric) fits perceived concepts involving multiple properties better than the degree 1 or city block distance measure appropriate for separable dimensions. In the geometric formulation, therefore, the conceptual space is a Cartesian product of domains rather than dimensions. Thus, as well as the distance metrics on the dimensions, the partitioning of dimensions into domains is a key element in the geometric representation.

The final key element of the geometric representation of conceptual space as described by Gärdenfors is the role of connectivity and convexity in properties and concepts, which have been captured in the definitions and results of Section 3. Connectivity is required in all regions, used in properties, concepts, and areas of attention. Prototypes have been used in transition rules, and their sensible definition relies on convexity.

Definition 6.2.1 (*Geometric conceptual space*). A formal definition of the conceptual space which captures the spirit of Gärdenfors' exposition is:

- $D = \prod_i \Delta_i$ where each Δ_i is a metric space—a perceptual quality—with a distinguished point at infinity, and the projection $\pi_i: D \rightarrow \Delta_i$ is the Cartesian projection.
- There is one symbol for each concept, including names of domains, dimension, and properties; denote the symbol dimensions L_j .
- Domains take the Euclidean metric formed from the metrics of the underlying dimensions D_i , and the relationship between the domains and the dimensions is determined by the partition $\{Y_i\}_{i \in Dom}$ of $Dim = \{1, 2, \dots, N\}$. Betweenness is induced by this metric.
- $v > 1$ and $s = 1$ (*required for categorisation*).

Let $i_i : D_i \rightarrow D$ be the canonical injection. Use this to identify D_i with a subspace in D . Recall that $P_{j,i}$ is the property in domain i associated with the j th symbol.

The type 1 and 2 transitions readily accommodate Gärdenfors’ Criterion C for a natural concept: i.e., as a region over domains which each have a salience weighting, together with information about how regions in different domains are correlated [15, p. 105]. The salience weighting is the importance function w used in the transition Eqs. (5.1) and (5.3), which in this setting become

$$c_i(t + 1) = \lambda \sum \{c(k, t)w(k, i) f(k, i): f(k, i) \in Att(0) \& c_i(t) = *\},$$

for $\lambda > 0$, (6.4)

$$c(k, t + 1) = c(k, t) + \sum_{ci(t) \in \pi_i(Att(0))} w(k, i) \exp(-d_i(c_i(t), f(k, i))).$$
(6.5)

A class of transitions is required to support induction. Here, we will not change the associations which define a concept—that part of learning is outside the scope of the present paper. Generalisation is only performed if the current state does not have a prior value in the generalisation domain, so that in our setting, data dominates rules. For example, a generalisation rule that all Frenchmen are good lovers will be blocked if Jacques is known to be a bad lover. Generalisation rules are related to type 1 transitions, in that they set a previously unknown value; however, they can be triggered without requiring symbols to be active.

Suppose the subject of generalisation is domain Δ_j , and J is the set of domains to be used as the basis of generalisation. Suppose further that $k(m)$ is the symbol associated with the property to be substituted. Then the generalisation rule is

$$c_j(t + 1) = f(k(m), j)$$

if $f(k(m), j) \in Att(0) \& c_j(t) = * \&$
 $(d(c_{r(i)}(t), f(k(m), r(i)))) < \delta$ for $r(i) \in J$.

(6.6)

This says that if $c(t)$ is close enough to some set of prototypical values in the domains indexed by J , but takes unknown value in Δ_j , which is in the current area of attention, then the current state should be set to the prototypical value for this domain for the supposed concept labelled by $k(m)$. Such a transition will lead to type 2 transitions on the symbol $k(m)$, which will raise the activity of the symbol of this property, thus notifying through the symbolic layer that the current state has been associated with the property. This in turn can be used in a categorisation problem using the type 1 and 2 transitions above.

We emphasise, however, that the output does not have to be “interpreted” by any higher level. Generalisation occurs as part of the processing, whenever the area of attention moves to the right domain and the current state for this domain is not defined. In particular, generalisation is seamless with the categorisation activity described in 5.3; the transitions occur if the preconditions are met, in the midst of type 1 and 2 transitions, and the generalised data is available for the categorisation process.

7. Conclusion

Differentiation and identification, the fundamental components of reasoning, must arguably be based on similarity judgements. Similarity and the related concept of distance are not naturally defined in connectionist and symbolic systems. Many disciplinary areas resort to feature space representations to define distance and similarity, but such feature spaces involve strong assumptions about the structure and independence of the feature dimensions. Moreover, in feature spaces names of structural elements are implicitly rather than explicitly linked.

Our work explores a new paradigm for representation and reasoning which overcomes some deficiencies in existing representations. It provides a formal and general investigation of conceptual spaces which captures the essential geometric elements of their development in Gärdenfors [15]. Conceptual spaces have been set in a mathematical category of metric spaces, which does not necessarily constrain them, but opens the way for geometric intuition based on our everyday experience of our 3-dimensional world, and provides access to machinery from algebraic geometry and functional analysis. Missing values and complex objects have been modelled. A symbol subspace has been introduced to support naming and communication, and to direct the dynamics of conceptual spaces. Properties have been defined in terms of the associations individuals make between labelled concepts and domains.

The existence of conceptual spaces as representations with the dynamics described here is not in question, since well known systems can be interpreted as conceptual spaces. This however leads to the question: what does the conceptual space formulation add to existing representation formulations? Certainly, metric concepts have been used for a long time in representation in many areas, particularly in computer vision and spatial modelling. They have been adjoined to symbolic systems, and to ANNs through control dynamics. Their use in conceptual spaces has been described at length in the work of Gärdenfors.

The contribution of this paper has been to provide a comprehensive formal foundation for a meso level representation in which the metric concepts are developed from the ground up, without assuming a Cartesian setting. Links to other forms of representation have been demonstrated. Dynamics have been developed which explain categorisation and convert nicely to dynamical system and neural network behaviours.

While this paper has added a formal foundation for conceptual spaces, it remains only an introduction. A great deal more work has to be done to reap the benefits of the meso level perspective on representation which, for developers of artificial cognitive systems, will derive from a representation and reasoning paradigm which naturally supports the notion of similarity judgements, but which has explicit links to ANNs and symbolic systems. Extensions will cover many disciplinary areas. They include tailored implementations of the dynamics, and practical application in domains such as medicine in which the metric representation would be linked to symbolic terminological systems. Theoretical amplification includes allowing the structuring of domains from dimensions to depend on the concept rather than to apply to all concepts, and using mappings between conceptual spaces to describe both learning and communication. The relationship between this formulation and other extensions of the symbolic and neural approaches should be further

explored. Laboratory experiments are needed to investigate convexity and compactness in conceptual representations, as well as to investigate non-Euclidean structure.

We have not required that the conceptual space be a vector space, which would allow addition and scalar multiplication. Nevertheless, there is cognitive support for the thesis that dimensions should be quantitative, with addition defined. To extend this to include multiplication over a field (most likely, the reals) may also be justified. Were conceptual spaces defined to be vector spaces, they would inherit the properties usually assumed of feature spaces in the huge amount of work on classification and pattern recognition carried out in many disciplinary areas. So a better understanding is needed of when assumptions about vector space structures and Minkowski metrics are valid in conceptual representation.

References

- [1] J. Aisbett, G. Gibbon, A tunable distance measure for coloured solid models, *Artificial Intelligence* 65 (1994) 143–164.
- [2] J. Aisbett, G. Gibbon, Conceptual spaces as voltage maps, *Proc. 6th International Work Conference on Artificial and Natural Neural Networks*, Granada, Spain, 2001.
- [3] L. Barsalou, Perceptual symbol systems, *Behavioral and Brain Sciences* 22 (1997) 577–660.
- [4] K. Borsuk, W. Szmielew, *Foundations of Geometry*, North-Holland, Amsterdam, 1960.
- [5] G. Carpenter, S. Grossberg, A massively parallel architecture for a self-organizing neural pattern recognition machine, *Computer Vision, Graphics, and Image Processing* 37 (1987) 54–115.
- [6] G. Carpenter, S. Grossberg, ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures, *Neural Networks* 3 (1990) 129–152.
- [7] G. Carpenter, S. Grossberg, J. Reynolds, ARTMAP: Supervised real-time learning and classification of non-stationary data by a self-organizing neural network, *Neural Networks* 4 (1991) 565–588.
- [8] A. Clark, *Sensory Qualities*, Clarendon Press, Oxford, 1993.
- [9] A. Chella, M. Frixione, S. Gaglio, Understanding dynamic scenes, *Artificial Intelligence* 123 (2000) 89–132.
- [10] S. Dehaene, Varieties of numerical abilities, *Cognition* 44 (1992) 1–42.
- [11] Z. Dienes, J. Perner, A theory of implicit and explicit knowledge, *Behavioral and Brain Sciences* 22 (1999) 735–808.
- [12] W. Freeman, Qualitative overview of population neurodynamics, *Neural Modeling and Neural Networks* (1994) 185–215.
- [13] P. Gärdenfors, A geometric model of concept formation, in: S. Ohsuga, H. Kangassalo, H. Jaakkola, K. Hori, N. Yonezaki (Eds.), *Information Modelling and Knowledge Bases III*, IOS Press, Amsterdam, 1992, pp. 1–16.
- [14] P. Gärdenfors, Symbolic, conceptual and subconceptual representations, in: V. Cantoni et al. (Eds.), *Human and Machine Perception: Information Fusion*, Plenum Press, New York, 1997, pp. 255–270.
- [15] P. Gärdenfors, *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, MA, 2000.
- [16] P. Gärdenfors, Concept combination: A geometrical model, in: L. Cavedon, P. Blackburn, N. Braisby, A. Shimojima (Eds.), *Logic Language and Computation*, Vol. 3, CSLI, Stanford, CA, 2000.
- [17] R. Goldstone, Influences of categorisation on perceptual discrimination, *J. Experimental Psychology: General* 123 (1994) 178–200.
- [18] N. Goodman, *Fact, Fiction, and Forecast*, Harvard University Press, Cambridge, MA, 1955.
- [19] C.L. Hardin, *Color for Philosophers: Unweaving the Rainbow*, Hackett, Indianapolis, IN, 1988.
- [20] S. Harnad, The symbol grounding problem, *Physica D* 42 (1990) 335–346.
- [21] C. Hempel, *Aspects of Scientific Explanation, and other Essays in the Philosophy of Science*, Free Press, New York, 1965.
- [22] M.S. Humphreys, J. Wiles, S. Dennis, Toward a theory of human memory: Data structures and access processes, *Behavioral and Brain Sciences* 17 (1994) 655–692.

- [23] G. Lakoff, *Women, Fire and Dangerous Things*, The University of Chicago Press, Chicago, IL, 1987.
- [24] D. Kirsch, Today the earwig, tomorrow man?, *Artificial Intelligence* 47 (1991) 161–184.
- [25] E. Margolis, S. Laurence (Eds.), *Concepts: Core Readings*, MIT Press, Cambridge, MA, 1999.
- [26] Markham, Dietrich, In defense of representation, *Cognitive Psychology* 40 (2000) 138–171.
- [27] R. Nosofsky, Similarity, scaling and cognitive process models, *Annual Review of Psychology* 43 (1992) 25–53.
- [28] A. Okabe, B. Boots, K. Sugihara, *Spatial Tessellations: Concepts and Applications*, Wiley, New York, 1992.
- [29] T. Omori, A. Mochizuki, K. Mizutani, M. Nishizaki, Emergence of symbolic behavior from brain like memory with dynamic attention, *Neural Networks* 12 (1999) 1157–1172.
- [30] M. Page, Connectionist modelling in psychology: A localist manifesto, *Behavioral and Brain Sciences* 23 (4) (2000) 443–467.
- [31] J. Paredis, The emergence of data structures from local interactions, in: Schwefel (Ed.), *Lecture Notes in Computer Science*, Vol. 496, Springer, Berlin, 1991.
- [32] D. Pisanelli, A. Gangemi, G. Steve, A medical ontology library that integrates the UMLS metathesaurus, in: *Proc. Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM 99*, 1999.
- [33] S. Quartz, T. Sejnowski et al., The neural basis of cognitive development: A constructivist manifesto, *Behavioral and Brain Sciences* 20 (4) (1997) 537–596.
- [34] W. Quine, Natural kinds, in: *Ontological Relativity and other Essays*, Columbia University Press, 1969, pp. 114–138.
- [35] Richer, A Practical Guide for Differentiating Between Iron Deficiency Anaemia and Anaemia of Chronic Disease in Children and Adults, *The Nurse Practitioner*, April 1997, <http://www.springnet.com/ce/j704a.htm>.
- [36] D. Roberson, J. Davidoff, N. Braisby, Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks, *Cognition* 71 (1999) 1–42.
- [37] R. Shepard, S. Chipman, Second order isomorphism of internal representation: Shapes of states, *Cognitive Psychology* 1 (1970) 1–17.
- [38] R. Shepard, Toward a universal law of generalization for psychological science, *Science* 237 (1987) 1317–1323.
- [39] R. Shepard, Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis, in: G. Lockhead, J. Pomerantz (Eds.), *The Perception of Structure*, APA, Washington, DC, 1991, pp. 53–72.
- [40] G. Simmons, *Introduction to Topology and Modern Analysis*, McGraw Hill, New York, 1963.
- [41] D. Tranel, H. Damasio, A. Damasio, A neural basis for the retrieval of conceptual knowledge, *Neuropsychologia* 35 (10) (1997) 1319–1327.
- [42] T. van Gelder, R. Port, It's about time. An overview of the dynamical approach to cognition, in: R.F. Port, T. van Gelder (Eds.), *Minds as Motion*, MIT Press, Cambridge, MA, 1995.
- [43] N.H. Williams, *Combinatorial Set Theory*, North-Holland, Amsterdam, 1977.