AI Magazine Volume 22 Number 1 (2001) (© AAAI)

Book Reviews

Conceptual Spaces—The Geometry of Thought

Norman Foo

Conceptual Spaces—The Geometry of Thought is a book by Peter Gärdenfors, professor of cognitive science at Lund University, Sweden. Gärdenfors has authored another book in this series (based on work with Carlos Alchourron and David Makinson), *Knowledge in Flux,* a definitive account of the widely examined AGM (after Alchourron, Gärdenfors, and Makinson) theory of belief revision. The AGM theory is firmly based on classical logic and its model theory, and by his founding participation in developing it, Gärdenfors has earned the right to critique knowledge representation. His new book is not primarily about logic, but it is certainly not an apostasy either. If I may be permitted a minor irreverence, I would say that this book came not to destroy logic but to fulfill.

Knowledge representation as we know it has reached an impasse. There seems to be two problems, or perhaps, they are merely different perspectives of an underlying problem. The first is knowledge acquisition, and the second is machine learning. *Conceptual Spaces* has important messages for both.

For machines to use knowledge it has to be somehow acquired. In principle, this knowledge can be hand coded into machines by experts or even elicited from them by an automated process. The practice is disappointing. The systems so constructed suffer from the aptly named brittleness problem, which is AI's version of the control engineer's sensitivity nightmare—small changes in the intended semantics of predicates or procedures cause large and usually catastrophic changes in performance. There have been, I think, two responses to this. One is owed to my colleague Paul Compton and his coworkers, using *ripple-down rules* (RDRs), which at the risk of oversimplification is the online adaptation of an expert system—users provide feedback as they apply the system in real life. The beauty of RDR is that the semantics is elastic, and in principle, the extension of any predicate is infinitely flexible. Moreover, users do not need to know anything about the mechanics of the system. The system seems to gradually acquire "community meanings" for predicates. However, there is a decided disadvantage. The declarative semantics of predicates is extremely unwieldy because it hinges on long chains of exceptions. The second response is to build ontologies, which has appeal because the fundamental idea is old and tested, witness Linneaus and botany. Moreover, the contemporary programming paradigm of object-orientation worships at the temple of ontologies. Whatever the reality might be, our knowledge seems to be organized as collections of classes that reside in trees with their implicit hierarchy. Much AI a decade or more ago was devoted to the algebras and logics of property and method inheritance among classes so ordered. Today, a lot of effort is going into the automation, or at least semiautomation, of the construction of such ontologies. There is as yet no convincing evidence that these two responses will succeed in large-scale practice.

Conceptual Spaces—The Geometry of Thought, Peter Gärdenfors, The MIT Press, Cambridge, Massachusetts, 2000, 307 pp., ISBN 0-262-07199-1.

Machine learning is commercially hot. There is no enterprise with access to large customer databases (social class, behavior patterns, credit history, and so on) that has not attempted to data mine them. My colleague Ross Quinlan is an eminent authority on the use of decision trees for data mining. Quinlan's book is now a standard reference for practitioners. Data mining is in fact an old idea if one were to understand it as a kind of systems identification. There is this amorphous pool of data that can be interpreted as output from an underlying system of rules and facts. The task is to describe this system. The devil is in the detail because the amorphous pool itself is in a sense its own best descriptor. Thus, a good data-mining description has to condense, summarize, and elucidate. The problem is that often this description can best be achieved by the introduction of abstract concepts or theoretical predicates, none of which present themselves as directly observable quantities in the databases. Familiar examples of such abstractions in the finance domain are concepts such as liquidity, velocity of money, and risk factor. The concepts that survive are those that resonate with user requirements, typically predictive or explanatory power. The invention of such abstractions is a creative task, and one of the aims of ontology research is to facilitate this creativity.

Well, can the creative process of concept invention be fully automated? In a strong sense, this book implicitly answers no because the semantics of concepts is dependent on context and purpose as well as the topological, algebraic, and logical structures used to represent them. Circumstance and teleology are as important as the formalism. However, the book provides valuable clues about how the creative process might be assisted by isolating some techniques that are promising from others that are probably deadends.

Gärdenfors addresses core issues in concept formation and concept structure in a novel setting, making connections to philosophy, psychology, and computer science. His book is therefore truly a monograph in cognitive science with many excursions into topics such as neural nets and nonmonotonic logics. One intriguing idea is the introduction of topology into concept structure, where the topology is determined by natural dimensions. Because topology is not usually part of existing knowledge bases (unless it is about spatial domains), perhaps a brief summary is in order. The topologies introduced in this book are characterized by regions that contain points. There are two properties of regions that are desirable: (1) connectedness and (2) convexity. A region is *connected* if it cannot be decomposed into two or more smaller nonintersecting regions. It is *convex* if every line that connects any two points passes only through the region. The notion of a line is contextual and, hence, so is that of convexity. Convexity can be quite subtle. To cite an example from the book, the color circle has sectors of an annulus (the region) representing a particular color, say, red (the closer to the center, the smaller the hue). If a straight line is used as a path between two points in this sector, sometimes it will fall outside the red sector; so, this sector is ostensibly not convex. However, the circle topology suggests that the correct coordinates to use are polar, not Cartesian. In polar coordinates, a *line* is an arc that is part of the circumference of a circle passing through the red sector; so, this sector is indeed convex. If the shape of the region is convex according to the underlying coordinates that can be used to describe it, then Gärdenfors makes a persuasive case that it corresponds to a natural concept.

A neat resolution of the Goodman paradox is provided as an example of the book's emphasis on convexity relative to the topologies as a requirement for natural concepts. It is argued that convexity sanctions the kinds of induction we find natural, and conversely, the lack of it is a cause of apparent paradox. The underlying topologies are often multidimensional, with each dimension corresponding to innate perceptual qualities. Examples of such dimensions are color (topology is the color circle), time (topology is the real line), and judgments of temperature (topology is an interval). A qualitative concept such as *hot* is convex relative to the temperature dimension—if an object is considered to be hot at temperatures $t1$ and $t2$, then it will also be considered hot at any temperature t such that $t1 < t < t2$ (which is the usual definition of convexity on real numbers). *Hot* is also a good example of a concept that is subject to what Gärdenfors calls *contextual effects,* for example, what might be hot for bath water is not necessarily hot for coffee. Prototype theory is another rich area in which the book's topological theme is explicated.

There is an enormous literature on concept classification that is reviewed in this book. Using its topological theme as the focus, interesting new light is shed on past work and results, including the classic Labov experiments on cup-bowl classifications, Voronoi diagrams, and Tversky's work on similarity measures. There is an illuminating discussion of concept learning as geometric dynamics in a metric space and a stringent criticism of existing nonmonotonic formalisms as sufficiently inexpressive because they ignore the underlying topological relations between concepts. If you are worried about the kind of strait jacket worn by strict belief revisionists who work within a given, unchanging language, Gärdenfors provides hints on how real ontology revision might work. It is important to bear in mind when reading this part of the book that the empirical details that seem so critical to any implementation would not make sense without the unifying thread of a conceptual space that runs through it.

Another intriguing idea is more like an implicit invitation to explore a proposal to solve the symbol-grounding problem. Put bluntly, the challenge is to account for how high-level concepts and their representational symbols can arise from low-level neural processes. In the penultimate chapter, the book theorizes that at least some of the account can be based on the distinction between slow versus fast dynamics of neural systems, where the former accounts for the emergence of concepts. This theory is plausible, for even in classical systems theory, this separation into slow and fast dynamics holds under loose assumptions, and over 30 years ago, research by the Nobel laureate Herbert Simon and his colleagues showed that aggregations of variables can be used as emergent symbolic descriptions of slow dynamics.

Although the book does not suggest how the knowledge representation impasse can be circumvented, it provides insights by telling us where to look for solutions and which proposals are bad prospects. It is essential reading for knowledge representation researchers and has a wealth of implicit research projects that will challenge the best cognitive scientists.

Norman Foo is professor of computer science and engineering and director of the Knowledge Systems Group in the AI Laboratory at the University of New South Wales, Sydney, Australia. He is a graduate of Canterbury University, New Zealand, and the University of Michigan at Ann Arbor. His current interests are in knowledge representation and reasoning, logic programming, and systems complexity. His URL is www.cse.unsw.edu.au/~norman.