Kirsh, D. Implicit and Explicit Representation, Encyclopedia of Cognitive Science, 2003

Implicit and Explicit Representation

Intermediate article

David Kirsh, University of California, San Diego, California, USA

CONTENTS

The degree to which information is encoded explicitly in a representation is related to the computational cost of recovering or using the information. Knowledge that is implicit in a system need not be represented at all, even implicitly, if the cost of recovering it is prohibitive.

INTRODUCTION

During the brief history of cognitive science, disputes have often arisen over how explicitly certain types of information (or knowledge) are represented in human cognitive systems, and what it means for information to be explicitly rather than implicitly represented in a system. Every few years it seems that new ways are discovered to build information into architecture, internal dynamics, and agent–environment interaction. Information that was stored in 'software' becomes integrated into 'hardware' in the next generation of systems, and then becomes integrated into a new 'architectural' design in the following generation. This seems to hold whether the software is made out of computer programming languages or malleable neural connections, whether the hardware is silicon or organic, and whether the architecture is a computer system or an anatomical plan that only changes across generations. When information is no longer encoded in distinct data structures with definite location and form, should we say that the information is still there, but represented in a more implicit form?

Questions about how information may be embedded in systems, how it may exercise a causal influence over processes, without being easily identifiable with a recognizable state, structure or process, make problematic the analysis of representation, and especially the analysis of what it means for information to be explicitly, as opposed to implicitly, represented in a cognitive system. There is a temptation to simply dismiss the matter as being of purely philosophical, or even semantic, interest. But the questions remain important because any hope of understanding how information enters into causal accounts of cognition must eventually explain the relation between the information attributed to a system, and appealed to in explanatory models, and the underlying states, structures or processes that serve as the vehicles for that information, and which are closer to the level of purely causal or neurophysiological explanation. Since there are many explanatory models in cognitive science that refer to information that is only implicitly present, it is important to clarify what implicit representation means and how it enters into underlying causal interactions.

EXPLICITNESS AND SYMBOL-PROCESSING MODELS

According to the symbol-processing view, all higher-order cognition, and most lower-order cognition too, is a computational process in which syntactically structured representations – such as sentences in an internal language of thought (Fodor, 1975; Pylyshyn, 1984), algebraic or graph structures (Chomsky, 1980; Minsky, 1975), or matrices of numbers (Marr, 1982) – are systematically transformed in a rule-driven or algorithmic manner. To understand, and therefore explain, a cognitive process, it is necessary to track the trajectory of informational states that the host system follows as it moves towards an explicit answer to a computational problem: for example, determining the meaning of an utterance, the implication of a previous thought, or the shape and visual appearance of an external object.

Since a cognitive system, on this view, is a mechanism for applying rules to structured

representations, it is natural to regard information as explicitly encoded if the structured state, process or form – the symbol structure – representing that information can be interpreted according to a well-behaved theory of content, such as a truth or model theory. We can then point to any given structured representation and say 'that form explicitly encodes this content'.

Although the data structures and representations involved in the various symbol-processing models (lattices, matrices, graph structures, extended first-order predicate calculus) go well beyond the structures we typically see in everyday life, this idea of 'implicit' and 'explicit' representation remains close to the everyday meaning of the terms used when discussing natural language. In ordinary parlance, we regard a fact to be 'explicitly' stated if it is expressed literally and unambiguously in a well-formed sequence of words, a sentence. Something that must be inferred from the sentence – because, for instance, it is presupposed, or because it is a consequence of the meaning of the words – is not explicitly stated; it is implicit in the meaning and context of utterance. Thus, when someone says 'when did you stop being a bachelor?' we take them to be explicitly asking a question about a date or time, but implicitly asserting that we no longer are a bachelor (the major presupposition), that we now have a wife (a consequence of the meaning of 'bachelor'), that we know when our marriage took place (a presupposition of the 'when' question), and so on.

On this account, information is 'explicitly' encoded if it can be read directly off a sentence without more inference than is required to understand the meaning of the words and their structure; while information is 'implicitly' encoded if additional inference or semantic processing is needed to recover it.

Using this intuitive analysis, Dienes and Perner (1999) have presented a theory of implicit and explicit knowledge that attempts to integrate and relate the divergent uses of the implicit–explicit distinction in different research areas, such as implicit and explicit memory, blindsight, automatic and controlled action, and development. They offer an explanation of why discussions of explicit and implicit knowledge are so often tied to concepts like consciousness, volitional control, and verbalizability.

Any theory about the nature of explicit and implicit representation based on our intuitions about what is explicit and implicit in natural language can be generalized to other declarative representations discussed in symbol-manipulation models: matrices, connected graphs, vectors, etc. But our intuitions about explicitness, even in natural language, are not complete. For instance, the sentence 'police police police police police' is grammatical and unambiguous. This sentence may be paraphrased as 'policemen who are policed by policemen also police policemen'. Since it has a unique syntactic and semantic identity it should qualify as an explicit representation. One need only understand the meaning of the words and their structure. Yet few people can actually recover this meaning because too much computation is involved in determining which sense of 'police' is being used in each position. Because of the computational complexity of interpreting the meaning of the sentence, the meaning is not easily recovered, it is not on the surface, and cannot be directly accessed. Consequently, we have conflicting intuitions about whether to say its meaning is explicitly encoded. On the one hand, it satisfies the basic truth-theory intuition 'here is the representation, there is the meaning', since it is well formed and unambiguous, so it is explicit on that criterion. On the other hand, few people can actually recover the information in the representation, so it ought be implicit on that criterion. Thus it seems that the intuitive theory is incomplete.

Such concerns about the computational processes involved in understanding the meaning of a representation suggest that a theory of implicit and explicit representation should be based more on the way representation and computation interrelate.

EXPLICITNESS AS A COMPUTATIONAL PROPERTY OF REPRESENTATIONS

One such theory involves measuring the computational effort required to extract, use, or interpret the information encoded in a representation (Kirsh, 1991). The computational complexity of the process of interpretation determines where on the continuum of explicit to implicit a given representation lies. If the interpretative process implements a constant-time algorithm, or extracts the content quickly and without substantial involvement of the rest of the cognitive system, then the information it extracts is directly available and hence explicitly encoded. For instance, the numeral '5' encodes the number 5 in English more explicitly than $\sqrt[5]{3125}$, even though both designate 5, because the value 5 can be read off directly from '5' by English speakers without performing lengthy computation.

It is assumed here that a representation is a well-defined state, structure or process, in a causal

system; that it encodes a specifiable informational content that can be harnessed by the causal system of which it is a part; and that it is possible to use techniques of computational complexity theory to measure the computational effort involved in recovering the information, providing we have a theory of the processes that use or extract that content.

So to determine how explicitly or implicitly a piece of information is represented, we need a substantive theory about the functioning of the different parts of a cognitive system. This must include how the system identifies the state, structure or process carrying the information, and how it exploits the information.

A plausible consequence of this approach is that individual capacities for memory, learning and other cognitive skills can affect how explicit a representation is. For one person, a certain representation may be identified and grasped directly. For another, the representation cannot be grasped without substantial computation. This conforms to both intuition and science. We now know a good deal about the computational costs associated with different neural-network methods of identifying and individuating states (structures or processes). It is known that there are computational trade-offs between the number of perceptron-like connections reaching out into a region where neural states encode information (the spatial complexity of the identifying process), and the amount of later processing required to correctly identify the state (the time complexity of the identifying process) (Minsky and Papert, 1969). Consequently, a neural state that in one person may be immediately identified – because identified in a highly parallel manner – and so meet the identifiability condition of explicitness, for another person may not be immediately identifiable – because, for example, this person has not yet developed efficient recognition methods, and so will take much longer to identify the information-bearing state and access the information contained therein. For a lengthy discussion of these points see Kirsh (1991).

IMPLICATIONS

By treating explicitness as a computational property of a representation, we have a method for deciding the point at which we can say that information or knowledge is so deeply embedded in hardware, architecture or agent–environment interactions that it is no longer a causally active agency and no longer represented, even implicitly. It is the point where the cost to the cognitive system of recovering the information would be so great that there are no connections or accessing procedures that can reliably make use of the information. Since a representation is a mechanism for reliably carrying information across space or time, it is best to say that the information is so implicitly built into the cognitive system as to no longer be represented in it. At the other end of the continuum, information that is encoded in well-defined states, structures or processes, which the cognitive system can immediately identify and use, is explicit.

A further consequence of tying the implicit–explicit distinction to computation is that we have a method for distinguishing different types of explanation in cognitive science. Not all explanations of behavior and design are mechanistic. We accept this in other design sciences, where discussions of design rationale and historical evolution help us to understand why a particular design is a good one. The same should hold true when we discuss the 'implicit theory of the world' built into a system or process.

Although sometimes such claims point the way to a computational explanation, we should not assume that every implicit theory implies that there are counterpart representations in the system. For instance, a vision module designed to extract a three-dimensional shape from two stereoscopic images works rapidly if it is equipped with an algorithm that differentiates. Such an algorithm will work if the assumption about the world that objects change in shape smoothly and continuously is true. Smoothness is a 'success condition' of the algorithm, and any designer who wishes to determine the algorithm's reliability will need to know how often the assumption is correct. But the algorithm may not implicitly represent the assumption. The assumption will not be recoverable by the system itself if the representational vocabulary of early vision does not include terms such as 'smoothness'. The assumption of smoothness serves as a powerful constraint on the design space of useful algorithms and representations; but it is not implicitly represented.

Thus, theoretical assumptions, such as smoothness, generativity, conformity to a truth theory, adherence to a formal linguistic theory, and so on, may guide the design of algorithms or the architecture of a system; they may figure in discussions of implicit knowledge or of the evolutionary fitness of a creature; but they need not figure directly in discussions of the computational mechanism. Depending on the cost of using those assumptions directly, they may be best understood as nonmechanistic explanations.

References

Chomsky N (1980) *Rules and Representations*. Oxford: Blackwell.

Dienes Z and Perner J (1999) A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences* **22**(5): 735–755.

Fodor JA (1975) *The Language of Thought*. New York, NY: Thomas Crowell.

Kirsh D (1991) When is information explicitly represented? In: Hanson PP (ed.) *Information, Language, and Cognition*, pp. 340–365. New York, NY: Oxford University Press.

Marr D (1982) *Vision*. New York, NY: W.H. Freeman.

Minsky M (1975) A framework for representing knowledge. In: Winston PH (ed.) *The Psychology of Computer Vision*, pp. 211–277. New York, NY: McGraw-Hill.

Minsky M and Papert S (1969) *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Pylyshyn ZW (1984) *Computation and Cognition*. Cambridge, MA: MIT Press.

Further Reading

Clark A (1993) *Associative Engines*. Cambridge, MA: Bradford Books.

Cummins R (1986) Inexplicit representation. In: Brand M and Harnish R (eds) *The Representation of Knowledge and Belief*. Tucson, AZ: University of Arizona Press.

Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske F (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Evans G (1982) *The Varieties of Reference*. Oxford: Oxford University Press.

Fodor JA (1981) *Representations*. Cambridge, MA: MIT Press.

Fodor JA and Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. In: Pinker S and Mehler J (eds) *Connections and Symbols*, pp. 3–71. Cambridge, MA: MIT Press.

Goodman N (1976) *Languages of Art*, 2nd edn. Indianapolis, IN: Hackett.

Palmer SE (1978) Fundamental aspects of cognitive representation. In: Rosch E and Lloyd B (eds) *Cognition and Categorization*, pp. 259–303. Mahwah, NJ: Erlbaum.

Implicit Cognition

Intermediate article

Eyal M Reingold, University of Toronto, Ontario, Canada
Colleen A Ray, University of Toronto, Ontario, Canada

CONTENTS

'Implicit cognition' refers to unconscious influences reflecting perception, memory, and learning, without subjective phenomenal awareness.

TERMINOLOGY AND DEFINITIONS

Since the 1970s, the 'implicit–explicit' distinction has become the dominant terminology under which experimental cognitive psychology has investigated unconscious influences on behavior and thought. Historically, the relation between consciousness and cognition has been one of the most controversial areas of investigation in psychology. This controversy can be traced back to the association of the conscious–unconscious distinction with psychoanalytic theory. Many experimental psychologists have regarded the unconscious as an unsuitable topic for scientific inquiry. But related phenomena have been empirically investigated under a variety of alternative terminologies. Terms such as 'incidental', 'pre-attentive', 'inaccessible' and 'covert' were used to avoid the unwanted associations of the terms 'unconscious' or 'unaware'. The term 'implicit' is a more recent, and perhaps more successful, example of such terminological camouflage. The implicit–explicit