Chapter 1

Introduction

Dan Sperber

Cognitive systems are characterized by their ability to construct and process mental representations. Cognitive systems capable of communicating also produce and interpret public representations. Representations, whether mental or public, are themselves objects in the world; they are found inside cognizers and in the vicinity of communicators; they are potential objects of second-order representations or "metarepresentations." While the term "metarepresentation" gained currency only in the late 1980s (early uses are found for instance in Pylyshyn, 1978; Sperber, 1985a; Leslie, 1987), the general idea is much older. Under a variety of other names, philosophers, psychologists, linguists, logicians, semioticians, literary theorists, theologians, and anthropologists have been interested in different types of metarepresentations. To give but one example, the seventeenth-century *Port-Royal Logic* devotes a chapter to the distinction between "ideas of things" and "ideas of signs," the latter being mental representations of public representations.

Mental representations of mental representations (e.g., the thought "John believes that it will rain"), mental representations of public representations (e.g., the thought "John said that it will rain"), public representations of mental representations (e.g., the utterance "John believes that it will rain"), and public representations of public representations (e.g., the utterance "John said that it will rain") are four main categories of metarepresentation. Most scholars have been predominantly interested in only one category: theory-of-mind psychologists, for instance, have studied mental representations of mental representations; reader-response theorists, mental representations of public representations; and semioticians, public representations of public representations.

Notwithstanding historical antecedents, much recent work on metarepresentations is truly novel as a result of being pursued within the framework of cognitive science and of philosophy of cognition. As such it gives great importance to mental representations of mental representations – drawing on relatively sophisticated notions of mental representation – whereas older work was mostly about the public representation of public representations – with elaborate taxonomies of public representations such as those found, for instance, in classical rhetoric or modern semiotics. Also, much current work is about the cognitive abilities that produce and exploit metarepresentations whereas older work was mostly about semi-formal properties of metarepresentations and about their role in communication. These earlier issues are still of great interest but their study is now approached from a cognitive rather than from a semiotic or hermeneutic point of view.

Reviewing past and present literature on metarepresentations would be a formidable task. Here, let me just briefly evoke four main areas where recent work on metarepresentations has been of particular importance: primate cognition, developmental psychology, philosophy of consciousness, and linguistic semantics and pragmatics.

While the ability to form representations is found in all animals with cognitive capacities, the ability to form metarepresentations is extremely rare. Most animal species, it is assumed, utterly lack metarepresentational abilities. In highly intelligent social animals such as primates, on the other hand, it has been argued that an ability to interpret and predict the behavior of others by recognizing their mental states may have evolved. In Dennett's terms (1987), some primates have been described as "second-order intentional systems," capable of having "beliefs and desires about beliefs and desires." Second-order intentional systems are, for instance, capable of deliberate deception. In a population of second-order intentional systems, a third-order intentional system would be at a real advantage, if only because it would be able to see through deception. Similarly, in a population of third-order intentional systems, a fourth-order intentional system would a greater advantage still, having greater abilities to deceive others and avoid being deceived itself, and so on. Hence, the hypothesis that an evolutionary arms race could have developed that resulted in a kind of "Machiavellian intelligence" consisting in higher-order metarepresentational abilities (Humphrey, 1976; Byrne & Whiten, 1988; Whiten & Byrne, 1997). Evolutionary and ethological arguments have sometimes converged with, sometimes diverged from, the experimental studies of primates' metarepresentational abilities that had started with Premack and Woodruff's pioneering article, "Does the chimpanzee have a theory of mind?" (1978).

Though the level of metarepresentational sophistication of other primates is still contentious, that of human beings is not. The human lineage may be the only one in which a true escalation of metarepresentational abilities has taken place.

Humans are all spontaneous psychologists. They attribute to one another many kinds of propositional attitudes: beliefs, regrets, opinions, desires, fears, intentions, and so on. Philosophers have described the basic tenets of this "folk psychology" and discussed its validity. Psychologists have focused on the individual development of this cognitive ability, often described as a "theory of mind." Philosophers and psychologists have been jointly involved in discussing the mechanism through which humans succeed in metarepresenting other people's thoughts and their own. This investigation has, in particular, taken the form of a debate between those who believe that attribution of mental states to others is done by means of simulation (e.g., Goldman, 1993; Gordon, 1986; Harris, 1989) and those who believe that it is done by deploying one kind or another of "theory" (e.g., Gopnik, 1993; Leslie, 1987; Perner, 1991, Wellman, 1990). In this debate between the simulation and the theory-theory views, much attention has been paid to different degrees of metarepresentational competence that may be involved in attributing mental states to others. In particular, the ability to attribute *false* beliefs has been seen as a sufficient, if not necessary, proof of basic metarepresentational competence. This metarepresentational competence can be impaired and this has been the basis of a new, cognitive approach to autism. Conversely, the study of autism has contributed to the development of a more fine-grained understanding of metarepresentations (see, Baron-Cohen, 1995; Frith, 1989; Happé 1994).

The ability to metarepresent one's own mental states plays an important role in consciousness and may even be seen as defining it. For David Rosenthal (1986, 1997), in particular, a mental state is conscious if it is represented in a higher-order thought. When a thought itself is conscious, then the higher-order thought that represents it is a straightforward metarepresentation. These higher-order thoughts may themselves be the object of thoughts of a yet higher order: the reflexive character of consciousness (i.e., the fact that one can be conscious of being conscious) is then explained in terms of a hierarchy of metarepresentations. (For other metarepresentational approaches to consciousness, see Carruthers 1996, Lycan 1996).

Cognitive approaches have stressed the metarepresentational complexity of human communication. It has been argued that the very act of communicating involves, on the part of the communicator and addressee, mutual metarepresentations of each other's mental states. In ordinary circumstances, the addressee of a speech act is interested in the linguistic meaning of the utterance only as a means to discover the speaker's meaning. Speaker's meaning has been analyzed by Paul Grice (1989) in terms of several layers of metarepresentational intentions, in particular the basic metarepresentational intention to cause in the addressee a certain mental state (e.g., a belief) and the higher-order metarepresentational intention to have that basic intention recognized by the addressee. Grice's analysis of metarepresentational intentions involved in communication has been discussed and developed by philosophers and linguists such as Bach & Harnish, 1979; Bennett, 1976;

Recanati, 1986; Schiffer, 1972; Searle, 1969; Sperber & Wilson, 1986. The metarepresentational complexity of human communication combines with that of language itself. It has long been observed that human languages have the semantic and syntactic resources to serve as their own metalanguage. In direct and indirect quotations, for instance, utterances and meanings are being metarepresented. The semantics of these metarepresentational uses has had a central issue in philosophy of language at least since Frege. Recent work in linguistics and pragmatics has suggested that there are several other, less obvious, metarepresentational dimensions of language use.

Under this name or another, a notion of metarepresentation has also been invoked in philosophical work on intentionality and on rationality (individual and collective), in the psychology of reasoning, and in epistemic logic, semantics, aesthetics, ethics, anthropology of religion, cognitive archeology, epistemology, philosophy of science, and artificial intelligence. This diversity of contributions is somewhat obscured by the extraordinary development of work in just one area: theory-of-mind (I am hyphenating the expression since it is often used without commitment to the view that the mind-reading ability in question is really theory-like; see, for instance, Alan Leslie, this volume). There have been, over the past 15 years, dozens of conferences, books, and special issues of journals devoted to the psychology of theory-of-mind (e.g. Baron-Cohen, Tager-Flusberg, & Cohen, 1993; Bogdan, 1997; Carruthers & Smith, 1996; Davies & Stone, 1995a; 1995b). In this literature, the link with autism, on the one hand, and with primate cognition, on the other, is often made. What is lacking is a much broader approach to metarepresentations, including theory-of-mind literature but not necessarily centered on it.

One could ask, though, whether there is a *general* story to be told about metarepresentations? Have works that make use of some notion of metarepresentation more in common than works that, in other domains, make use of, say, some notion of 'symmetry', or of 'reward'? Is there any good reason to have them confront one another? I have come to think that the answer to these questions is "yes."

In my own work, I have used a notion of metarepresentation in research on cultural symbolism (1975), on apparently irrational beliefs (1982/1985; 1997), on anthropological hermeneutics (1985b), on the evolution of language (1994), on the dynamics of culture (1985a; 1996) and, with Deirdre Wilson (Sperber & Wilson, 1986/1995), on communicative intentions, on irony and metaphor (Sperber & Wilson, 1981; 1990, Wilson & Sperber, 1992), on speech acts (Wilson & Sperber, 1988), and on higher-level explicatures (Wilson & Sperber, 1993). I have found it more and more illuminating to think of all these metarepresentational phenomena as based on a metarepresentational capacity no less fundamental than the faculty for language. Understanding the character and the role of this metarepresentational capacity might change our view of what it is to be human. This sentiment was reinforced by the interdisciplinary seminar on metarepresentations organized by Gloria Origgi at the CREA in Paris between 1994 and 1996, where many more metarepresentational issues were discussed, in particular with Daniel Andler, Pierre Jacob, and François Recanati. So, when Steven Davis invited me to organize the Tenth Vancouver Cognitive Science Conference, I had little hesitation: the conference would be on metarepresentations and would definitely bring together participants who had explored the notion in quite different ways.

The conference took place at Simon Fraser University in Vancouver, Canada, in February 1997. The present volume is a collection of essays based on the talks given at the conference and revised in the light of our debates. The chapters are organized in three parts: (1) The evolution of metarepresentation, (2) Metarepresentations in mind, and (3) Metarepresentations, language, and meaning. While this organization reflects three dominant themes of the conference, there is an unavoidable degree of arbitrariness in assigning individual chapters to parts. Several chapters are relevant to more than one major theme and other themes – the rudimentary forms of metarepresentation, the contrast between internalist and externalist views of representations, or the metarepresentational character of suppositional thinking – link chapters in yet other ways. Here is a brief guide to the contents of the volume.

The Evolution of Metarepresentation

In "Making tools for thinking," **Daniel Dennett** raises fundamental challenges. The notion of a metarepresentation cannot be clearer than that of a representation. The notion of a representation can be understood in a variety of senses, some shallower and wider, such that we would be willing to attribute representations to simpler animals and devices. Other senses are narrower and richer, such that we might be tempted to think of representation as specifically human. Do these richer senses of "representation" somehow presuppose that representations are being (or are capable of being) metarepresented? Can we conceive of the emergence in evolution and in cognitive development of metarepresentations – and of the type of representations that requires metarepresentation – in a way that is purely internal to the mind or should we see this emergence as linked to the availability in the environment of representational tools – linguistic symbols, for instance – that are there to be metarepresented? These issues are well worth keeping in mind when reading the rest of the book

In "The mind beyond itself," **Robert Wilson** speculates on issues similar to those raised by Dennett. He criticizes the individualistic approach to cognition and develops the idea that that many higher cognitive functions and, in particular, metarepresentational capacities are essentially world-involving. He discusses the cases of memory, theory-of-mind, and cultural evolution and argues that, in each case, external symbols and their metarepresentations play an essential role.

In "Consider the source: The evolution of adaptations for decoupling and metarepresentations," **Leda Cosmides and John Tooby** outline a novel and wide-ranging approach to the evolution of metarepresentational abilities. They start from the observation that human evolution is characterized by a dramatic increase in the use of contingent information for the regulation of improvised behavior tailored to local conditions. They argue that adaptations evolved to solve the problems posed by using local and contingent information include a specialized "scope syntax," decoupling systems, and a variety of metarepresentational devices. These adaptations are essential to planning, communication, mind-reading, pretence, deception, inference about past or hidden causal relations, mental simulation, and much else. Thus Cosmides and Tooby view mind-reading as only one of the functions that has driven the evolution of metarepresentational abilities and of human intelligence in general. One may note that the representational powers they see as having evolved in the human mind are interestingly similar to those François Recanati analyzes from a semantic point of view in his chapter.

In "Metarepresentations in an evolutionary perspective," **Dan Sperber** envisages the possibility that humans might be endowed, not with one, but with several evolved metarepresentational abilities. He argues that, beside the standard metapsychological mind-reading ability, humans might have a comprehension module aimed at the on-line interpretation of utterances, and a logico-argumentative module, aimed at persuading others and avoiding deception.

In "Chimpanzee cognition and the question of mental re-representation," **Andrew Whiten** examines the state of the evidence regarding the ability of chimpanzees to engage in imitation, mind-reading, and pretence. He argues that chimpanzees have a capacity for a most basic form of metarepresentation, which he calls "re-representation." These are mental representations whose content derives from other mental representations either in oneself or in others. He discusses how these abilities in apes relate to the different "grades" of metarepresentation envisaged in the theory-of-mind literature, in particular by A. M. Leslie and J. Perner. This chapter provides an appropriate transition to the second part.

Metarepresentations in Mind

In "The mentalizing folk," **Alvin Goldman** raises central questions regarding people's abilities to metarepresent mental representations. What concepts of mental states do people possess? How do they attribute specific instances of mental states to themselves and to others? How do these abilities develop? He reviews the main competing answers to these questions, criticizes various forms of the theory-theory approach, and defends a version of the simulation-theory approach where particular attention is paid to introspection.

In "How to acquire a representational theory of mind," **Alan Leslie** discusses several versions of the theory-theory of cognitive development in its application to the acquisition a representational theory-of-mind. Theory-theories associate the possession of a concept – in particular, the concept of belief – to some descriptive knowledge of the referents, in this case, of beliefs. Leslie argues against this view and for a "conceptual psycho-physical" approach where a concept such as that of belief might be causally correlated with, or "locked to," beliefs in the world and be that concept just because of this locking mechanism. The concept of belief, then, is not acquired as part of a proper "theory" of mind. Rather, the acquisition of a theory is made possible by the possession and deployment of the previously available concept. What makes this concept of belief available – as well as the basic metarepresentational abilities where it gets deployed – may well be an innate disposition rather than a learning process.

In "Metarepresentation and conceptual change: Evidence from Williams Syndrome," **Susan Carey and Susan Johnson** present a case study of abnormal cognitive development, specifically, the acquisition of a intuitive but non-core theory of biology by a population of retarded people with Williams Syndrome. They argue that the bootstrapping devices that underlie conceptual change require metarepresentational cognitive architecture. Metarepresentational capacities that are part of the theory-of-mind module support, for instance, noticing of contradictions and distinguishing appearance from reality, thus permitting conceptual change. However, in the case of retarded individuals, the lack of sufficient computational capacity serves as a bottleneck both in the construction of metaconceptual knowledge that goes beyond the core and in the construction of the first theories that likewise transcend the core. This study also throws light on the status of the four-year-old's theory-of-mind as core knowledge or constructed knowledge.

David Rosenthal's HOT (i.e., higher-order thought) theory of consciousness is a particularly clear and crisp case of metarepresentational thinking. In "Consciousness and metacognition," he defends this theory and discusses relevant evidence from current research on metacognition and, in particular, on feeling-of-knowing experiences. He argues that this evidence sheds light on what it is to be conscious of a mental state and on what it is, therefore, for a mental state to be conscious. He discusses important issues having to do with the development of metacognitive abilities and with their fallibility.

Metarepresentations, Language, and Meaning

In "Meaning, exemplarization and metarepresentation," **Keith Lehrer** argues that the human mind is essentially a "metamind" (see Lehrer, 1990), involving first-level representational states that are metarepresented and evaluated at a metalevel, thus becoming states of the metamind. This permits mental plasticity and the resolution of conflicts that, at the lower level, are unavoidable for a complex representational system. Such a metarepresentational view seems, however, threatened by a regress (as suggested by Wilfrid Sellars) or by circularity (as suggested by Jerry Fodor) in accounting for language learning. Drawing on Sellars theory of meaning, and on Nelson Goodman's notion of exemplarization, Lehrer argues that the problem of understanding meaning and of achieving representational transparency is resolved through a harmless referential loop of ascent to quotation and descent to disquotation.

In "The iconicity of metarepresentations," **François Recanati** develops an extensive and original formal treatment of the semantics of metarepresentations. He discusses the relevant philosophical literature on quotations and indirect reports of speech or thought and argues, against standard views, for a Principle of Iconicity according to which true metarepresentations essentially resemble the representations they are about. They are fundamentally "transparent," in that they represent what the metarepresented representation represents and not just, "opaquely," that representation itself. He contrast his approach to the simulation view of metarepresentations and speculates about the relationship between conditionals and metarepresentations.

In a series of influential papers, Tyler Burge has argued for the view that the intentional states of a subject are, in part, determined by the social practices of the members of his community. The disposition to defer to experts plays an important role in this externalist view. In "Social externalism and deference," **Steven Davis** discusses and refines Burge's account. He argues that a conditional disposition to defer is essential to the possession of concepts. He analyzes this disposition to defer as involving epistemic norms and a metarepresentational ability. This chapter thus relates the metarepresentational framework to some of the most interesting recent developments in the philosophy of language and mind.

In "Metarepresentations in staged communicative acts," **Raymond Gibbs** demonstrates, with linguistic and experimental evidence, how speakers' and listeners' recognition of specific metarepresentations affects their joint production and understanding of nonserious speech and, in particular, irony. The evidence tends to show that irony, because of its complex metarepresentational character, requires more processing effort to understand than tropes like metaphor. Gibbs concludes that the most general challenge that studying metarepresentations in language poses is to recognize how the coordination of mutual beliefs in ordinary speech reflects essential connections between the ways people think and the ways they produce and understand language.

In "Metarepresentation in linguistic communication," **Deirdre Wilson** examines the different types of metarepresentational ability involved in linguistic comprehension. She discusses Grice's metarepresentational view of speaker's meaning and of processes of comprehension. Focusing on the use of utterances to represent attributed utterances and thoughts, she surveys a range of linguistic metarepresentational devices and argues that their analysis can both benefit from, and provide useful evidence for, the study of more general metarepresentational abilities. From an historical point of view, current approaches to metarepresentations derive from semiotic and philosophical interest in metalinguistic devices. Deirdre Wilson's chapter, showing how this traditional interest is now being reframed in a cognitive perspective, provides a fitting conclusion for the whole volume.

References

Bach, Kent, & Harnish, Robert (1979). *Linguistic communication and speech acts.* Cambridge, MA.: Harvard University Press.

Baron-Cohen, Simon (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press.

Baron-Cohen, Simon., Leslie, Alan, & Frith, Uta (1985). Does the autistic child have a "theory of mind"? *Cognition 21*, 37–46.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J., Eds. (1993). *Understanding other minds: Perspectives from autism.* Oxford: Oxford University Press.

Bennett, J. (1976), *Linguistic behaviour.* Cambridge: Cambridge University Press.

Bogdan, R. J. (1997). *Interpreting minds: The evolution of a practice.* Cambridge, MA: MIT Press.

Byrne, R. W., & Whiten, A. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes and humans.* Oxford: Oxford University Press.

Carruthers, P. (1996). *Language, thought and consciousness.* Cambridge: Cambridge University Press.

Carruthers, P., & Smith, P., Eds. (1996). *Theories of theories of mind.* Cambridge: Cambridge University Press.

Davies, M., & Stone, T., Eds. (1995a). *Folk psychology: The theory of mind debate.* Oxford: Blackwell.

Davies, M., & Stone, T., Eds. (1995b). *Mental simulation: Evaluations and applications.* Oxford: Blackwell.

Dennett, D. (1987). *The intentional stance.* Cambridge: MIT Press.

Frith, U. (1989). *Autism: Explaining the enigma.* Oxford: Blackwell.

Gibbs, R. (1994). *The poetics of mind: Figurative thought, language and understanding.* Cambridge: Cambridge University Press.

Goldman, A. (1993). The psychology of folk psychology. *The Behavioral and Brain Sciences 16,* 15–28.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *The Behavioral and Brain Sciences 16,* 1–14.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language 1,* 158–171.

Grice, H. P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Happé, F. (1994). *Autism: An introduction to psychological theory.* London: UCL Press.

Harris, P. L. (1989). *Children and emotion: The development of psychological understanding.* Oxford: Blackwell.

Humphrey, Nicholas K. (1976). The social function of the intellect. In P. P. G. Bateson and R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303–317). Cambridge: Cambridge University Press.

Lehrer, K. (1990). *Metamind.* Oxford: Oxford University Press.

Leslie, A. M. (1987). Pretence and representation: The origins of "theory of mind." *Psychological Review 94,* 412–426.

Lycan, William. (1996). *Consciousness and experience.* Cambridge, MA: MIT Press.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Science 1,* 515–526.

Pylyshyn, Zenon W. (1978). When is attribution of beliefs justified? *The Behavioral and Brain Sciences 1,* 592–526.

Recanati, F. (1986). On defining communicative intentions. *Mind and Language 1* (3), 213–242.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies* 49 (3), 329–359.

Rosenthal, D.M. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds), *The nature of consciousness: Philosophical debates* (pp. 729–753). Cambridge, MA: MIT Press.

Schiffer, S. (1972). *Meaning.* Oxford: Clarendon Press,

Searle, J. (1969). *Speech acts.* Cambridge: Cambridge University Press.

Sperber, Dan (1975). *Rethinking symbolism.* Cambridge: Cambridge University Press.

Sperber, Dan (1982/1985). Apparently irrational beliefs. In S. Lukes & M. Hollis (Eds.), *Rationality and relativism* (pp. 149–180). Oxford: Blackwell. Revised edition in Dan Sperber (Ed.), *On anthropological knowledge.* Cambridge: Cambridge University Press.

Sperber, Dan (1985a). Anthropology and psychology: Towards an epidemiology of representations. (The Malinowski Memorial Lecture 1984).*Man (N.S.) 20,* 73–89.

Sperber, Dan (1985b). *On anthropological knowledge.* Cambridge: Cambridge University Press.

Sperber, Dan (1994). Understanding verbal understanding. In J. Khalfa (Ed.), *What is intelligence?* (pp. 179–198). Cambridge: Cambridge University Press.

Sperber, Dan (1996). *Explaining culture: A naturalistic approach.* Oxford: Blackwell.

Sperber, Dan (1997). Intuitive and reflective beliefs.*Mind and Language 12,* 67–83.

Sperber, Dan, & Wilson, Deirdre (1981). Irony and the use-mention distinction. In P. Cole (Ed.), *Radical pragmatics* (pp. 295–318). New York: Academic Press.

Sperber, Dan, & Wilson, Deirdre (1986/1995). *Relevance: Communication and cognition.* Oxford: Blackwell. Second Edition 1995.

Sperber, Dan, & Wilson, Deirdre (1990). Rhetoric and relevance. In John Bender & David Wellbery (Eds.), *The ends of rhetoric: History, theory, practice* (pp. 140–156). Stanford, CA: Stanford University Press.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Whiten, A. (1991). *Natural theories of mind: Evolution, development and simulation of everyday mind-reading.* Oxford: Basil Blackwell.

Whiten, A. & Byrne, R. W. (1997). *Machiavellian intelligence II: Evaluations and extensions.* Cambridge: Cambridge University Press.

Wilson, Deirdre, & Sperber, Dan (1988). Mood and the analysis of non-declarative sentences. In Jonathan Dancy, Julius Moravcsik, & Charles Taylor (Eds.), *Human agency: Language, duty and value* (pp. 77–101). Stanford, CA: Stanford University Press.

Wilson, Deirdre, & Sperber, Dan (1992). On verbal irony. *Lingua 87,* 53–76.

Wilson, Deirdre, & Sperber, Dan (1993). Linguistic form and relevance. *Lingua 90,* 1–25.