Chapter 5

Metarepresentations in an Evolutionary Perspective

Dan Sperber

Just as bats are unique in their ability to use echo-location, so humans are unique in their ability to use metarepresentations. Other primates may have some rather rudimentary metarepresentational capacities. We humans are massive users of metarepresentations and of quite complex ones at that. We have no difficulty, for instance, in processing a three-tiered metarepresentation such as that in example (1).

(1) Peter thinks that Mary said that it is implausible that pigs fly.

The fact that humans are expert users of metarepresentations, is, I would argue, as important in understanding human behavior as the fact that bats are expert users of echo-location is in understanding bat behavior.

How has the human metarepresentational capacity evolved? In order to contribute to the ongoing debate on this question, I will focus on three more specific issues:

(i) How do humans metarepresent representations?

(ii) Which came first, language or metarepresentations?

(iii) Do humans have more than one metarepresentational ability?

How Do Humans Metarepresent Representations?

Metarepresentations are representations of representations but not all representations of representations are metarepresentations in the relevant sense. The human metarepresentational capacity we are interested in here is, first and foremost, a capacity to represent the *content* of representations. Consider:

(2) (a) Bill had a thought.

(b) This claim is often repeated.

Statements (2a) and (2b) are about representations but they do not in any way represent their contents. They are not metarepresentations in any useful sense. One can imagine a biological or artificial device that could detect the presence of representations but not at all their content properties. It could detect, say, mental representations by being sensitive to the appropriate manifestations of brain activity or it could detect public representations such as utterances by being sensitive to their phonetic properties. Such a device, lacking access to the content properties of the representations it would represent, would not have a metarepresentational capacity in the sense intended. Consider:

(3) (a) Mary is hypochondriac.

(b) Henry is always complaining.

(c) John is a creationist.

(d) This claim is slanderous.

Statements (3a) through (3d) do attribute a more or less specific tenor to representations they are directly or indirectly about. Thus, (3a) is true only if Mary tends to form unwarranted beliefs the gist of which is that she is seriously ill; (3b) is true only if Henry is making claims to the effect that certain things are not the way they ought to be; (3c) is true only if John accepts the central tenets of the creationist doctrine; (3d) is true only if the claim, to which reference is made, attributes to some agent some undesirable property. Representations (3a) through (3d) do state or imply something about the content of some representations, although they do not articulate the contents of these representations. They are metarepresentational only in a rudimentary way.

There may be animals capable of detecting the presence and tenor of some mental representations in others but who do so by means of unstructured representations. Thus, an animal might detect the fact that a conspecific wants to mate and represent this by means of a single unarticulated symbol *wants-to-mate* as in statement (4a). Similarly an animal might detect the fact that a conspecific has been suitably impressed with some display of strength and represent this by means of a single unarticulated symbol *knows-that-I-am-stronger-than-him* as in statement (4b).

(4) (a) He wants-to-mate.

(b) He knows-that-I-am-stronger-than-him.

Such animals would possess a very rudimentary metarepresentational capacity lacking compositionality and recursion. They could only metarepresent a short and fixed list of representations.

Imagine a species with quite limited cognitive capacities. The contents of its mental representations belong to a limited repertoire: "there is a predator (at a certain location)"; "there is a prey"; "there is a mating partner," and so forth. Suppose it were advantageous for members of this species to be capable of representing some of the mental states of their conspecifics. Then, a rudimentary metarepresentational ability of the kind that I have just evoked might evolve.

Consider now a species with rich cognitive capacities, capable of forming and of using in its behavior mental representations of indefinitely varied contents. Suppose again that it were advantageous for members of this species to be capable of representing indefinitely many of the mental states of their conspecifics. A rudimentary metarepresentational ability of the kind just considered would not do this time since it would be sufficient to represent only a narrow and fixed subset of the indefinitely varied mental representations of the members of the species.

Another way to put the same point is this. Consider first the case of a species with a system of internal representations consisting, roughly, of a small list of representation types that can occasionally be tokened (and indexed to the situation). Then, the internal representations of members of this species could be metarepresented by means of a rudimentary metarepresentational system with a different mental symbol for each metarepresentable representation type. Consider now a species with system of internal representations that is – or is equivalent to – a mental language with compositionality and recursion. Then, the internal representations of members of this species could be metarepresented only by means of metarepresentational system consisting in – or equivalent to – a meta-language no less rich than the language the expressions of which it serves to metarepresent. If, moreover, several levels of metarepresentation are possible, as in statement (1) above, the metarepresentational system must equally rich or richer at each level. The only cost-effective way to achieve this is to have the expressions of the object-language do double service as expression of the meta-language (or, if *n* levels of metarepresentation are possible, do *n+1*-tuple service). A full-fledged metarepresentational capability such as that found in human languages and in human thinking is based on the possibility of interpreting any expression-token as representing another token of the same expression or the expression-type or, more generally, some expression type or token it resembles in relevant respects. Thus, in examples (5a) through (5e) the expressions italicized represent not the state of affairs that they describe but representations (mental, public, or abstract) the contents of which they serve to render.

(5) (a) Bill thought that *the house was on fire.*

(b) The claim that $2 + 2 = 4$ is often repeated.

(c) Mary believes that *she is seriously ill.*

(d) Henry is complaining that *life is too short.*

(d) John believes that *God created the world.*

(e) The claim that *John is a creationist* is slanderous.

Examples (1) and (5a) through (5e) are linguistic but I take it that humans are capable of entertaining their mental equivalent. In other words, expressions in the internal system of conceptual representations (however close or distant from natural language this system might be) can serve to represent expression types or tokens that they resemble in relevant respects (and identity is best seen here as just a limiting case of resemblance; see, Sperber & Wilson, 1995, chap. 4).

Imagine an organism endowed with a rich internal system of conceptual representations but without the ability to use these "opaquely" or metarepresentationally, that is, as iconic representations of other representations (types or tokens). Could such an organism *learn* to do so? Think of what is involved. The organism has no more knowledge of its internal representations *qua* representations than it has of its patterns of neural activities. Its representations are wholly transparent to it. Such an organism might be capable of representing, say, the fact that it is raining but never of representing the fact that it is representing the fact that it is raining.

It is hard to imagine what combination of external stimulations and internal dispositions might ever cause individual organisms of this type to become capable of using their repertoire of mental representations in order to represent not what these representations transparently represent but, in virtue of resemblance relationships, to represent opaquely other representations. This seems as implausible as learning to guide one's spontaneous movement by means of echo-location in the absence of a genetically determined domain-specific disposition. Such abilities speak of biological and evolutionary rather than cognitive and developmental transitions. I am, of course, aware of the relative weakness of "hard-to-imagine-otherwise" arguments. Still, if you are arguing that a full-fledged metarepresentational ability is something learnt (in the sense of learning theory) and do not illuminate how this learning might take place, then you are signing a huge promissory note the equivalent of which has never been honored. If, on the other hand, you assume that a metarepresentational ability is a biological development, the promissory note you are signing is more modest: there are well-understood biological developments of much greater complexity.

Which Came First: Language or Metarepresentations?

Language – rather than metarepresentational ability – is usually taken to be the most distinctive feature of the human species. The two are clearly linked as natural languages serve as their own meta-language and thus incorporate a full-fledged metarepresentational capacity. Linguistic utterances are public representations and typical objects of mental metarepresentation. Speakers, in intending an utterance, and hearers, in interpreting an utterance, mentally represent it as a bearer of specified content, that is, they metarepresent it.

Language and metarepresentations are made possible by biologically evolved mental mechanisms, which, it has been argued in both cases, are domain-specific. Noam Chomsky arguing for a domain-specific language faculty (e.g., Chomsky 1975, 1980) introduced the very idea of domain-specificity to the cognitive sciences. The idea of metarepresentations became familiar in cognitive science through work on naïve psychology or "theory of mind." In suggesting, in their article, "Does the chimpanzee have a theory of mind?" that the ability to attribute mental states to others was also found among non-linguistic animals, Premack and Woodruff were implying that this ability, at least in its simpler forms, is independent of language (Premack and Woodruff, 1978; see also, Premack, 1988). Developmental psychologists have argued that theory of mind is based on a domain-specific mental module (Leslie 1987; 1994; Baron-Cohen 1995).

If one accepts, as I do, the existence of two dedicated mental mechanisms, one for language, the other for metarepresentations, it seems reasonable to assume that, in humans, they have co-evolved. While the fully developed version of each of these two mechanisms may presuppose the development of the other, it still makes sense to ask which of these two, the linguistic or the metarepresentational, might have developed first to a degree sufficient to bootstrap the co-evolutionary process. At first blush, the language-first hypothesis seems quite attractive – and has attracted, for instance, Daniel Dennett (1991).

The hypothesis that the language faculty evolved first may seem, moreover, to offer a way of explaining how a metarepresentational ability might emerge in individual cognitive development, even in the absence of a biologically evolved specialized disposition. Linguistic communication fills the environment with a new kind of object – utterances, that is, public representations. Utterances can be perceived, attended to, thought about, just as any other perceptible object in the environment. At the same time, they are representations, they have meaning, content. It may seem imaginable, then, that children, finding linguistic representations in their environment, grasp the representational character of these utterances

because they are linguistically equipped to assign them content and, as a result, develop an ability to represent representations *qua* representations. This acquired metarepresentational ability would apply first to linguistic utterances and, then, would extend to other types of representations, in particular mental representations.

This hypothesis loses much of its appeal under scrutiny. Spelling it out results in a worrisome dilemma. On the one hand, we can argue that ancestral linguistic communication, though presumably simpler in many respects, was based on the same kind of communicative mechanism as modern linguistic communication. If so, it presupposed metarepresentational ability and, therefore, could not precede it. On the other hand, we can argue that ancestral linguistic communication was strictly a coding-decoding affair like other forms of non-human animal communication. There is then no reason to assume that our ancestors had the resources to become aware of the representational character of their signals anymore than bees or vervet monkeys do.

When we, modern humans, communicate verbally, we decode what the words mean in order to find out what the speaker meant. Discovering the meaning of a sentence is just a means to an end. Our true interest is in the speaker's meaning. A speaker's meaning is a mental representation entertained by the speaker that she intends the hearer to recognize and to which she intends him to take some specific attitude (e.g., accept as true). Verbal understanding consists in forming a metarepresentation of a representation of the speaker (in fact, a higher-order metarepresentation since the speaker's representation is itself a metarepresentational intention). Moreover, there is a systematic gap between the sentence's meaning and the speaker's meaning. Sentences are typically ambiguous and must be disambiguated; they contain referring expressions the intended referent of which must be identified; they underdetermine the speaker's meaning in many other ways. Linguistic utterances fall short, typically by a wide margin, of encoding their speaker's meanings. On the other hand, utterances are generally good pieces of evidence of these meanings. Inference to the best explanation of the speaker's linguistic behavior generally consists in attributing to her the intention to convey what actually was her meaning in producing her utterance. Linguistic comprehension is an inferential task using decoded material as evidence. The inferences involved are about the speaker's meaning, that is, they are aimed at metarepresentational conclusions.

If the ability to communicate linguistically had preceded the ability to use metarepresentations, then this pre-metarepresentational, ancestral verbal ability would have been radically different from the kind of verbal ability we modern humans use, which is metarepresentational through and through. The ancestral language would have been a coding-decoding affair as are the many forms of non-human animal

communication of which we know. This, in itself, is an unattractive speculation since it implies a radical change in the mechanism of human linguistic communication at some point in its evolution.

Even more importantly, there is no reason to assume that a decoding animal experiences or conceptualizes the stimulus it decodes as a representation: for non-human animals, coded communication and attribution of mental states to others are two unrelated capacities (the second apparently much rarer than the first). What seems to happen, rather, is that the decoding of the stimulus automatically puts the animal in a specific cognitive, emotional, or motivational state appropriate to the situation. For instance, the decoding of an alarm signal typically triggers a state of fear and activates escape plans. It would be quite unparsimonious to hypothesize that the decoding animal is able to recognize the signal as a signal endowed with meaning. It is more sensible (and the burden of proof would be on whoever maintained otherwise) to assume that, to communicators who merely code and decode, signals are transparent. Such signals play a role comparable to that of proximal stimuli in perception; they occur at an uncognized stage of a cognitive process.

If our ancestors were such coders-decoders and had no evolved disposition to metarepresent, then there is no sensible story of how the presence of utterances in their environment would have led them to discover their representational character, to metarepresent their content, and to use for this their own mental representations in a novel, opaque, manner. Out goes the hypothesis that language developed first.

What about the hypothesis that metarepresentations developed first? A metarepresentational and, more specifically, a metapsychological ability may be advantageous and may have evolved on its own. This has been convincingly argued in much recent literature on "Machiavellian intelligence" (Byrne & Whiten, 1988; Whiten & Byrne, 1997). The ability to interpret the behavior of intelligent conspecifics not just as bodily movement but as action guided by beliefs and desires gives one a much-enhanced predictive power. Predicting the behavior of others helps to protect oneself from them, to compete successfully with them, to exploit them, or to co-operate more profitably with them. A metarepresentational ability is plausible as an adaptation quite independently of communication.

Moreover, a well-developed metarepresentational ability makes certain forms of communication possible quite independently from any code or language. Organisms with metarepresentational abilities live in a world where there are not only physical facts but also mental facts. An individual may form beliefs or desires by emulating those it attributes to another individual. An individual may want to modify the beliefs and desires of others. It may want others to become aware of its beliefs and desires and to emulate these. Let me illustrate. Imagine two of our hominid ancestors, call one Mary and the other Peter.

First scenario
Mary is picking berries. Peter happens to be watching Mary. He infers from her behavior that she believes that these berries are edible and, since he assumes she is knowledgeable, he comes to believe that they are. Peter is using his metarepresentational ability to form new beliefs not just about Mary's mental representations but also about the state of affairs Mary's representations are about. He comes to "share" a belief of Mary's. Mary, however, is unaware that Peter is watching her and she has no desire to affect his beliefs. Peter in this case has a first-order metarepresentational belief:

Mary believes

that these berries are edible

Second scenario
Mary is aware that Peter is watching her and that he is likely to infer from her behavior that the berries are edible, and she intends him to draw this inference. Her behavior has now two goals: collecting berries and affecting Peter's beliefs. Peter, however, is unaware of Mary's intention to affect his beliefs. In an interestingly different scenario, Mary could believe that the berries are *in*edible and pick them in order to deceive Peter. In either case, Mary has a first-order metarepresentational intention:

That Peter should believe

that these berries are edible!

Third scenario
Peter is aware that Mary is picking berries with the intention that he should come to believe that these berries are edible. Mary, however, is unaware of Peter's awareness of her intention. How should Peter's awareness of Mary's intention affect his willingness to believe that the berries are edible (and to fulfil, thereby, Mary's intention)? If he believes that she is trying to be helpful to him by informing him that the berries are edible, this will give him extra reason to accept that they are. If, on the other hand, he mistrusts her, being aware of her informative intention will be a reason not to fulfil it. In either case, Peter has a second-order metarepresentational belief:

Mary intends

that he should believe

that these berries are edible.

Fourth scenario
Mary intends that Peter should be aware of her intention to inform him that the berries are edible. She has, then, not one but two informative intentions: a first-order informative intention that Peter should believe that the berries are edible and a second-order informative intention that Peter should be aware of her first-order informative intention. What reasons might she have to have the second-order informative intention? As mentioned above, Peter's awareness of Mary's first-order informative intention, provided he trusts her, may give him an extra reason to believe her. In other words, the fulfillment of the second-order informative intention may contribute to the fulfillment of the first-order informative intention. The second-order informative intention is, of course, a third-order metarepresentation to the effect:

That Peter should believe

that Mary intends

that he should believe

that these berries are edible!

Fifth scenario
Peter is aware that Mary intends him to be aware of her informative intention. He has a fourth-order metarepresentational belief:

Mary intends

that he should believe

that she intends

that he should believe

that these berries are edible.

Peter might come to have this belief when he notes that Mary is ostensively making sure that he is paying attention to her behavior by, say, establishing eye contact with him, picking the berries in somewhat formal manner, and so forth. Mary's first-order informative intention is now an "overt" (Strawson, 1964) or "mutually manifest" (Sperber & Wilson, 1995). We have reached a level where communication proper occurs, though no code or language is involved.

Once the level of metarepresentational sophistication of our fifth scenario is reached, a dramatic change indeed occurs. Whereas before,

in order to fulfil her first-level informative intention, Mary had to engage in behavior – picking the berries – that was best explained by attributing to her the belief that the berries were edible and the desire to collect the berries, she can now resort to symbolic behavior, the best explanation of which is simply that she is trying to fulfil an informative intention, her desire to inform (or misinform) Peter that the berries are edible. Instead of actually picking the berries, she might, for instance, mime the action of eating the berries.

Typically, symbolic behavior such as miming has no plausible explanation other than that it is intended to affect the beliefs or desires of an audience. This generally is the one effect of such a behavior that could clearly have been intended by the agent. Thus. for Mary to mime eating the berries does not feed her, does not help her in any easily imaginable way, except through the effect it has on Peter. Her miming behavior triggers in Peter's mind, through the perception of a resemblance between the miming behavior and the behavior mimed, the idea of eating the berries. Moreover, if it is at all relevant to Peter to know whether or not the berries are edible, Mary's behavior suggests that they are. Provided that Peter sees this miming behavior as intended by Mary to inform him of an informative intention of hers, he is justified in assuming that the very idea triggered in his mind was one Mary wanted him to entertain, elaborate and accept.

The same result achieved by miming could, provided Peter and Mary shared an appropriate code, be achieved by means of a coded signal or by means of some combination of iconic and coded behavior. Suppose Mary and Peter shared a signal meaning something like "good." Mary might point to the berries and produce this signal, again triggering in Peter's mind the idea that the berries were good to eat, doing so in an manifestly intentional manner and, therefore, justifying Peter in assuming that she intended him to believe that the berries were edible.

Note that, if Mary used the coded symbol "good" in this way, she would, nevertheless, be very far from *encoding* her meaning that these berries are edible. She would merely be giving evidence of her intention to cause Peter to come to accept this meaning as true. The use of coded signals as part of the process of communication is no proof that the communication in question is wholly, or even essentially, of the coding-decoding kind.

Metarepresentational sophistication allows a form of inferential communication independent of the possession of a common code. This type of inferential communication, however, can take advantage of a code. It can do so even if the signals generated by the code are ambiguous, incomplete, and context-dependent (all of which linguistic utterances actually are). By triggering mental representations in the audience, coded signals provide just the sort of evidence of the communicator's

informative intention that inferential communication requires, even if they come quite short of encoding the communicator's meaning.

To conclude this section, there is a plausible scenario where a metarepresentational ability develops in the ancestral species for reasons having to do with competition, exploitation, and co-operation and not with communication *per se*. This metarepresentational ability makes a form of inferential communication possible initially as a side effect and, probably, rather painstakingly at first. The beneficial character of this side effect turns it into a function of metarepresentations and creates a favorable environment for the evolution of a new adaptation, a linguistic ability. Once this linguistic ability develops, a co-evolutionary mutual enhancement of both abilities is easy enough to imagine.

Do Humans Have More than One Metarepresentational Ability?

Current discussions have focused on the metapsychological use of a metarepresentational ability to represent mental states, the beliefs and desires of others and of oneself. Humans, however, metarepresent not just mental representations but also public representations and representations considered in the abstract, independently of their mental or public instantiation. The three types of metarepresented representations are simultaneously illustrated in (1) and individually highlighted in (6a) through (6c).

(1) Peter thinks that Mary said that it is implausible that pigs fly.

(6) (a) Peter thinks that *Mary said that it is implausible that pigs fly.*

(b) Mary said that *it is implausible that pigs fly.*

(c) It is implausible that *pigs fly.*

In example (6a), the italicized phrase metarepresents a mental representation of Peter's, in (6b), it metarepresents an utterance, that is, a public representation of Mary's, and in (6c), it metarepresents a representation considered in the abstract (as a hypothesis), independently of whoever might entertain or express it.

Representations considered in the abstract are reduced to their logical, semantic, and epistemic properties: they may be true or false, self-contradictory or necessarily true, plausible or implausible, standing in relationships of entailment, of contradiction, of warrant, of being a good argument one for the other, of meaning similarity, and so forth. They may be normatively evaluated from a logico-semantic point of view (and also from an aesthetic point of view).

Mental and public representations have the same content properties as their abstract counterparts – or arguably, abstract representations are nothing but the content properties of concrete representations, abstracted away. Mental and public representations also have specific properties linked to their mode of instantiation. A mental representation occurs in one individual; it is causally linked to other mental and non-mental states and processes of which the individual is wholly or partially the locus. A mental representation can be sad or happy, disturbing or helpful; it can be normatively evaluated in psychological terms as poorly or well reasoned, as imaginative, as delusional, and so on. A public representation typically occurs in the common environment of two or more people; it is an artifact aimed at communication. It exhibits such properties as having a certain linguistic form, as being used to convey a certain content in a certain context, as attracting more or less attention, as being more or less comprehensible to its intended audience, and so on. It can be normatively evaluated from a communicative point of view as sincere or insincere, intelligible or unintelligible, relevant or irrelevant.

The properties of these three types of representations – mental, public, and abstract – do not constitute three disjoint sets and are not always easy to distinguish. Mental and public representations have the logico-semantic properties of their abstract counterparts. A belief or an utterance is said, for instance, to be true or false when its propositional content is. Public representations serve to convey mental representations and have, at least by extension, some of the properties of the mental representations they convey. An utterance is said, for instance, to be imaginative or delusional when the thought it expresses is. Still, many properties are best understood as belonging essentially to one of the three types of representation and as belonging to another type of representation, if at all, only by extension, in virtue of some relational property that holds between them (such as the relationship of expression that holds between an utterance and a thought).

There is no question that we modern humans can attend in different ways to these three types of representations. We can attribute mental states, interpret public representations, and reflect on the formal properties of abstract representations. Are these performances all based on a single metarepresentational ability or do they, in fact, involve different competencies? In the latter case, is it plausible that these competencies might each be a distinct evolved adaptation? Could there be several metarepresentational "modules"? In the literature of evolutionary psychology, on the one hand, and in the literature of developmental psychology, on the other, the only metarepresentational adaptation envisaged is a metapsychological "theory of mind," the main function of which is to predict the behavior of others. Even a peculiar behavior such

as collective pretend play, which involves essentially the creation and manipulation of public representations, is often treated as a regular use of "theory of mind."

What does a metarepresentational ability whose function is basically metapsychological do? What kind of inferences does it draw? Which properties of representations does it attend to? It draws inferences from situation and behavior to mental states as in examples (7a) through (7c), from mental states to other mental states as in examples (8a) and (8b), and from mental states to behavior as in examples (9a) and (9b).

(7) (a) There is a predator just in front of A.
Therefore, A knows that there is a predator just in front of it.

(b) A is panting.
Therefore, A wants to drink.

(c) A is running, occasionally looking behind its back.
Therefore, A is trying to escape.

(8) (a) A knows that there is a predator just in front of it.
Therefore, A is afraid of the predator.

(b) A wants to drink.
Therefore, A wants to find some water.

(9) (a) A is afraid of the predator.
Therefore, A will try to escape.

(b) A wants to find some water.
Therefore, A will go to the river.

A metapsychological ability assumes that others have some basic knowledge and basic drives, and attributes to them specific beliefs and desires derived through perception and inference. Does such an ability, with the kind of power we are entitled to grant it on evolutionary and developmental grounds, suffice to explain the kind of metarepresentational processes we engage in when we interpret public representations, in particular utterances, or when we attend to the formal properties of abstract representations?

A comprehension module?

Comprehension (or its pragmatic layer) is an inferential process, using as input the output of linguistic decoding and aiming at discovering the speaker's meaning. Comprehension consists, therefore, in inferring a mental state (an intention of a specific kind) from behavior (an utterance). It might seem that this is precisely the kind of result a metapsychological ability should be able to achieve. In the story above of Mary, Peter, and the berries, I tried to illustrate how, in principle, a multi-

leveled metapsychological ability might make inferential communication possible. The communication achieved between Mary and Peter in the fifth scenario showed metarepresentational prowess. In modern humans, however, comparable or greater metarepresentational achievements have become routine and we communicate at great speed much more complex contents than Mary and Peter ever could. This is a first reason to hypothesize that a more specialized adaptation aimed at comprehension has evolved. A second reason has to do with the pattern of inference from observed behavior to attributed intention.

In the simplest case, behavior (say, throwing a stone) is observed to have several effects (making a noise, casting a moving shadow, killing a bird). One of these effects (killing a bird, as it might be) can be seen as desirable to the agent. It is then inferred that the agent performed the behavior with the intention of achieving this desirable effect. Inferring intentions is somewhat harder when the behavior fails to achieve its intended effect. Say the bird is merely frightened away rather than killed. Competent attributers of mental states will nevertheless recognize that throwing the stone could have killed the bird and that the agent could expect this effect (with greater or lesser confidence). They will then infer that this unachieved effect was the one intended.

In the case of attributions of a speaker's meaning, these standard patterns of inference (from behavior to intention through identification of a desirable actual or failed effect) are not readily available. The essential effect intended by a speaker, that is, comprehension of her meaning, cannot be achieved without the very recognition of her intention to achieve this effect. The intended effect cannot, therefore, be independently observed to occur and then be recognized as desirable and presumably intentional. Nor does it make sense to imagine that the comprehender might recognize as the effect intended an effect that might plausibly have occurred but that, in fact, failed to do so, since the recognition of the intended effect would secure its occurrence.

There are, in the literature, two basic ways of solving the puzzle raised by inferential comprehension. The first way is Grice's (1989). It aims at showing a way around the difficulty while remaining within the limits of standard belief-desire psychology. It consists in assuming that the decoding of the linguistic signal provides a default assumption regarding the speaker's meaning. By default, the speaker is assumed to mean what the sentence she utters means. This default assumption can be inferentially complemented or otherwise corrected when there is a mismatch between it and general assumptions about standing goals that the speaker is presumed to aim at in her verbal behavior, goals codified by Grice in terms of his Co-operative Principle and Maxims. Such inferential correction involves a form of metarepresentational reasoning about the speaker's intentions in which not only the conclusion (an

attribution of meaning to the speaker) but also some of the premises are metarepresentational. For instance, in the case of a metaphor such as example (10), the hearer is supposed to reason somewhat as in (11):

(10) John is a soldier.

(11) (a) The speaker seems to have said that John is a soldier.

(b) The speaker does not believe and knows that I know that he does not believe that John is a soldier.

(c) The speaker is respecting the Co-operative Principle and, in particular, is trying to be truthful.

(d) Therefore, the speaker could not mean that John is a soldier.

(e) The speaker must be trying to convey a closely related meaning compatible with the presumption that the speaker is co-operative.

(f) By inference to the best available explanation, the speaker means that John is like a soldier: he is devoted to his duty, obedient to orders, and so on.

It is as if, by flouting the maxim of truthfulness, the speaker deliberately failed to achieve a certain effect, thus suggesting that the truly intended effect is in the vicinity of the overtly failed one.

As this example illustrates, Gricean pragmatics can be seen as an account where the inferential part of comprehension consists in applying common-sense psychology to verbal behavior. If Grice were right, a general metapsychological ability, together with a presumably socially acquired knowledge of the Co-operative Principle and the Maxims, would be sufficient to account for inferential comprehension.

Gricean pragmatics might seem attractively parsimonious since it does not require any specialized comprehension ability. The economy in terms of the number of mental devices one may be led to postulate, however, is more than offset by the cumbersome character of the inferences that Gricean pragmatics necessitates every time a speaker's meaning diverges from sentence's meaning (and we have argued that it *always* so diverges). Do we really have, in the case of implicature, of indirect speech acts, of metaphor, or of irony, to reflect on what the speaker knows we know she knows, on what respecting the maxims requires of her, on what she might mean and not mean? Does it take more effort and hence longer to process such utterances? (The answer is no; see Gibbs, 1994; this volume). Do we want to attribute to young children these complex inference patterns or to deny them the ability to comprehend metaphor and other forms of so-called indirect speech? As a rational reconstruction of how inferential comprehension might be possible, Grice's

account, though not without problems, is certainly appealing. As a psychologically realistic account of the mental processes actually involved in comprehension, it is much less so.

Though we owe a great deal to Grice's inspiration, Deirdre Wilson and I have criticized his pragmatics and, among other aspects, this account of metaphor. We have developed relevance theory as another way to solve the puzzle raised by inferential comprehension (Sperber and Wilson, 1995). In relevance theory, we assume that comprehension follows a very specific inferential pattern suited for the discovery of the informative intentions of communicators. We define the relevance of a cognitive input to an individual as a positive function of the cognitive effects achieved by processing this input and as a negative function of the amount of effort involved in this processing. We argue that every utterance (and, more generally, every communicative act) conveys a presumption of its own relevance. We show that this justifies the following inferential procedure: follow a route of least effort in constructing an interpretation – taking the very fact that an element of interpretation comes first to mind as an argument in its favor – until the effects achieved are sufficient to warrant the presumption of relevance conveyed by the utterance, and then stop. To illustrate briefly: given the context, utterance (10) might be capable of activating in the mind of the hearer the ideas in (11a) through (11g) in that order.

(10) John is a soldier.

(11) (a) John is devoted to his duty.

(b) John willingly follows orders.

(c) John does not question authority.

(d) John makes his own the goals of his team.

(e) John is a patriot.

(f) John earns a soldier's pay.

(g) John is a member of the military.

Suppose that the utterance, when interpreted as conveying (11a) though (11d), satisfies the expectations of relevance it has itself raised. Then, we predict that the interpretation will stop at this point and that (11g), that is, the literal interpretation will not even be considered. In another context, the order in which the elements of interpretation might come to the mind of the hearer would be different and the stopping point might be such that the overall interpretation would include (11g) and be literal. Suppose, for instance, that "John is a soldier" was said in response to the question "What does John do for a living?" Then (11a through (11g)

would probably be accessible in the reverse order, from (11g) through (11a). The hearer, moreover, would likely stop at (11f) or (11e). The resulting interpretation (11g) through (11e) would be literal and not even overlap with the metaphorical interpretations (11a) through (11d) of the same sentence uttered in a different context. In all cases, however, the comprehension procedure is the same: follow the path of least effort until adequate relevance is achieved. This may yield a literal, a loose, or a metaphorical interpretation without the comprehender having to take notice of the type of interpretation achieved.

The *conclusion* of such a process of interpretation is an attribution of a meaning to the speaker and, hence, a metarepresentation. Nevertheless, the *premises* in the inference process need not be metarepresentational. This procedure, therefore, can be followed by a relatively unsophisticated metarepresenter, for instance by a young child capable, as young children are, of comprehending metaphor. On the other hand, how would the child or, for that matter, the adult discover this procedure and recognize that it is a reliable means to infer a speaker's meaning? A plausible answer is that this procedure is not individually discovered but is biologically evolved. It is an evolved module.

The relevance-based comprehension procedure could not be soundly applied to the discovery of non-communicative intentions. Non-communicative behavior carries no presumption of relevance to possible observers. The order in which elements of interpretation come to the mind of the observer has no particular epistemic value. There is no level where, expectations of relevance being satisfied, the observer is thereby justified in believing that his interpretation of the agent's intention is complete.

What could be the relationship between a relevance-based comprehension module and a more general metapsychological module? The former might be a sub-module of the latter, in the manner in which linguistic acoustic abilities are a sub-module of general acoustic abilities. Speech sounds are sounds and their perception recruits the hearing system. At the same time, speech sounds are, from birth onwards, attended and processed in a proprietary way. Similarly, the recognition of communicative intentions might be a biologically differentiated and stabilized sub-system of human naïve psychology.

A "Logical" Module?

Is the human metapsychological ability enough to explain the human ability to attend to formal and, in particular, to logical properties of representations? A metapsychological ability attributes inferences to others. Inferences must, on the whole, respect logico-semantic relationships such as entailment or warrant – that is, relationships holding among representations in the abstract – or else they are not truly inferences. Successful

attribution of inferences to others must also respect these relationships. Still, if we suppose that the organisms to which mental states are attributed have quite limited inferential abilities, then the attributing organisms – the metapsychologists – need attend only to a few logical relationships. Moreover, this attention need not be reflective. That is, the metapsychologists need not be logicians, thinking about logical relationships; it is enough that they themselves be able to make, off-line so to speak, the inferences they attribute. (This is sometimes presented as the simulation view but there is a weak and sufficient version of it that is equally compatible with the simulation view, *à la* Goldman and with the theory-of-mind view, *à la* Leslie.)

The type of metapsychological competence that is likely to be an evolved adaptation is unlikely to explain the modern human ability to attend to abstract representations and, in particular, to do formal logic. However, it is dubious that such an ability is widely shared. It might be just a skill acquired with some difficulty by a minority of people in scholarly institutions.

In fact, psychologists favorable to an evolutionary point of view have expressed doubts as to the existence of a genetically determined domain-general "logical ability." If by a "logical ability" what is meant is a unitary, domain-general ability that would govern all human inferences, then indeed, positing such an ability would go against well-known evolutionary arguments for the domain-specificity of mental mechanisms (see Sperber, 1996, chap. 5). On the other hand, if we think of the kind of logical ability that is involved in formulating or evaluating arguments, far from being domain-general, this is a highly domain-specific metarepresentational ability: its domain is some specific properties of abstract representations. An appearance of domain-generality might come from the fact that the representations that such a device would handle could be representations of anything. The device, however, would just attend to some formal properties of these representations, independent of what they happen to represent, and would be as specialized as, say a grammar-checker in a word-processor, which is able to process statements about anything but is obviously not domain-general for all that.

Still, what is the plausibility, from an evolutionary point of view, that a logical module specialized in checking the validity of arguments would have evolved? It might seem low but let me speculate and offer an argument why such a logical module with a fairly specific function is not so implausible.

Though the conclusion is quite different, the structure of my argument is parallel to Leda Cosmides's argument on cheater detection (Cosmides, 1989). She draws on the neo-Darwinian argument according to which "reciprocal altruism" is unstable unless there is some form of

control for cheating since, otherwise, cheaters would be more successful until there were not enough genuine altruists left to be cheated. In the case of complex forms of reciprocal exchange, as found among humans, the prevention of cheating is an elaborate cognitive task that is likely, Cosmides argues, to have caused the evolution of an ad hoc adaptation, a cheater-detection mechanism. As usual, it is an open question as to whether the adaptation that handles the problem is precisely calibrated for the task or is, in fact, a narrower or, on the contrary, a larger ability that was still the best biologically available solution for the task. (I do not want to discuss Cosmides's basic argument, which I find illuminating, but to draw inspiration from it; for a critical discussion of her use of the selection task to provide empirical evidence for this argument, see Sperber, Cara & Girotto, 1995).

Communication is a form of co-operation that seems particularly advantageous for animals that depend as much as humans do on their cognitive resources. Instead of being restricted in one's knowledge to the products of one's own experiences and thinking, communication makes experience and thinking available by proxy. Alas, as with other forms of co-operation, communication makes one also vulnerable to misinformation, deception, and misguidance. Of course, one could protect oneself from the harms of deception by being systematically mistrustful but one would lose, by the same token, the benefits of communication. Communication is advantageous only if it is paired with mechanisms that ensure the proper calibration of trust. It is even more advantageous if, while protected from the deception of others without being overprotected, you can penetrate their protection and deceive them.

The human reliance on communication is so great, the risks of deception and manipulation so ubiquitous, that it is reasonable to speculate that all available cost-effective modes of defense are likely to have evolved. There are several such modes. One is to be discriminating about whom to trust and to accept authority quite selectively. Clearly, humans do this. Another is to be sensitive to subtle signs of deceptive intent, to read the relevant attitudinal and emotional signs. These safeguards can be breached, as they are by professional swindlers who typically "look absolutely trustworthy."

I want to focus on yet another possible protective mechanism against misinformation: check the consistency of the information communicated – both its internal consistency and its consistency with what you already believe. As anybody who has ever tried to lie knows, the liar's problem is to maintain consistency, both internally and contextually. An organism that drew its knowledge only from its senses and a few reliable inferential routines would probably waste energy in checking the consistency of the information it acquired. Inconsistencies in perception-based information occur but they are rare, and trying to eliminate

them may not be worth the cost. For richly communicating animals like us, eliminating inconsistencies may not only be worth the cost but, indeed, literally life-saving.

For this reason, I hypothesize the emergence of an ability to check for consistency and to filter incoming information on this basis. But, this is only the first blow in the persuasion counter-persuasion arm race. More advantageous than merely protecting yourself from misinformation is to combine this with the ability freely to inform and misinform others, to persuade them of what it is in your interest to persuade them of, whether true or false. Persuaders addressing consistency-checkers cannot do better than display the very consistency – or, at least, the appearance of it – for which their audience is likely to check. Communicators now express not only the propositions they want their audience to accept but also arguments to accept these propositions and the very argumentative structure that leads to the intended conclusion. The language is enriched with logical terms ("and," "or," "if," etc.) and para-logical terms ("therefore," "but," "since," "although," "even," etc.) that display the real or alleged consistency and logical force of the communication.

The next stage in the persuasion counter-persuasion arm race is the development of the ability to scrutinize these argumentative displays and to find fault with them. In other terms, I am surmising, on evolutionary grounds, the development of a very special kind of "logical," or logico-rhetorical ability. It is special in that it attends to logical relationships, not *per se* nor for the sake of the benefits that good reasoning can give the individual thinker, but, on the one hand, as a means to filter communicated information and, on the other hand, as a means to penetrate the filters of others. Such a speculation has experimentally testable implications. It predicts, for instance, that logically equivalent tasks will yield significantly better performance when they are presented in a context where subjects are scrutinizing arguments plausibly aimed at persuading them than when they are evaluating these arguments in the abstract (as happens in most experimental tasks).

Conclusion

My aim in the present chapter has been to discuss the possibility that humans might be endowed, not with one, but with several evolved metarepresentational abilities. I have considered argument as to why, beside the standard metapsychological ability, they might have a comprehension module aimed at the on-line interpretation of utterances and a logico-argumentative module aimed at persuading others while not being too easily persuaded themselves. If there are, indeed, three such metarepresentational modules, the obvious next step would be to envis-

age their evolution not separately but, much more systematically than I have done here, as a case of co-evolution involving also the "language instinct," and – dare I say it? – consciousness.

Acknowledgments

I am grateful to Steven Davis, Gloria Origgi, and Deirdre Wilson for their useful comments on earlier versions of this chapter.

References

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

Byrne, Richard, & Whiten, Andrew, Eds. (1988). *Machiavellian intelligence.* Oxford: Clarendon Press.

Chomsky, Noam (1975). *Reflections on language.* New York: Random House.

Chomsky, Noam (1980). *Rules and representations.* New York: Columbia University Press.

Cosmides, Leda (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with Wason Selection Task. *Cognition 31,* 187–276.

Dennett, Daniel (1991). *Consciousness explained.* Boston: Little, Brown.

Gibbs, Raymond (1994). *The poetics of mind.* Cambridge : Cambridge University Press.

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

Leslie, Alan (1987). Pretense and representation: The origin of "theory of mind." *Psychological Review 94,* 412–426.

Leslie, Alan (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). New York: Cambridge University Press.

Premack, D (1988). "Does the chimpanzee have a theory of mind?" revisited. In Richard Byrne and Andrew Whiten (Eds.), *Machiavellian intelligence* (pp. 161–179). Oxford: Clarendon Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences 1,* 532–526.

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

Sperber, Dan, Cara, Francesco, & Girotto, Vittorio (1995). Relevance theory explains the selection task. *Cognition 57,* 31–95.

Sperber, Dan, & Wilson, Deirdre (1995). *Relevance: Communication and cognition.* Oxford: Blackwell. (Original work published in 1986.)

Strawson, R. (1964). Intention and convention in speech acts. *Philosophical Review 73,* 439–460.

Whiten, Andrew, & Byrne, Richard (1997). *Machiavellian intelligence II: Extensions and evaluations.* Cambridge: Cambridge University Press.

This page intentionally left blank