Intentionality: Natural and Artificial

Colin Allen
Department of Philosophy
Texas A&M University
College Station, TX 77843-4237 USA
email: colin.allen@tamu.edu

February 26, 2001

Abstract

What role should philosophical theories of intentionality play in cognitive science? Philosophers argue on conceptual grounds about the appropriateness or inappropriateness of attributing intentional states to animals, to computers, and even to humans. I argue that a comparative approach to cognitive science (CACS) allows cognitive scientists to sidestep these disputes by treating philosophical theories as starting points for empirical investigation. For this purpose, philosophical theories of intentionality are most useful when they are naturalistic and not anthropocentric, and when they provide a framework for empirical research that does not place *a priori* constraints on what kinds of things can possess intentionality. I will apply these criteria critically to recent philosophical views of intentionality and I will show how particular attention to the comparative part of CACS should influence attempts to develop a theoretically useful conception of intentionality. In particular, this perspective shows that insistence on defining intentionality prior to experimental work is misguided; rather, an empirical, comparative approach to refining the notion of intentionality should be attempted. Finally, I discuss the consequences of different conceptions of intentionality for empirical work in ethology and artificial intelligence.

1 Introduction

Our common-sense theory of human minds involves the attribution of mental states that have “content”. For example, to have a belief one must have a belief *about* something and to have a desire one must have a desire *for* something. Philosophers label such states ‘intentional’. The term ‘intentional’ and its cognates have both an ordinary sense and a philosophical sense. The ordinary sense of ‘intentional’ connotes purpose and it is the sense in which *actions* can be said to be intentional. In this paper, however, the term is used in its *philosophical* sense, which is concerned with the idea that certain mental states are representational, or “about” other states of affairs. Intentionality is not limited to folk-psychological notions like belief and desire. Notions like mental representation and information which are widely used by cognitive scientists are also intentional in the philosophical sense (Allen 1992a; Allen & Hauser 1993).

Philosophers do not agree about the correct way to define intentionality. Nor do they agree that it can be defined in a way that makes questions about the mind empirically tractable. If intentionality cannot be treated empirically, then much of our common sense view of minds will not find a place in cognitive science (see, e.g., Churchland 1986), so this topic is of utmost importance to cognitive science. The gridlock caused by philosophical controversies about intentionality often causes empirical scientists to turn away from philosophical theorizing about these issues, but I will attempt to show that this would be a mistake, and that cognitive science can proceed without precise definition. I will not presuppose familiarity with the philosophical literature on intentionality so the next section contains a brief introduction. For the remainder of this section it will be adequate to think of intentionality as “aboutness” in the way that, e.g., one’s thoughts can be about Provence.

Comparative approaches to cognitive science (CACS) provide the best hope for relieving the philosophical gridlock surrounding intentionality. Both philosophers and cognitive scientists stand to benefit from the perspectives that arise when one attempts to compare cognitive abilities across species and outside the biological realm. CACS will involve the synthesis of perspectives from humans (cognitive psychology), non-human organisms (cognitive ethology), and computers (artificial intelligence). Foundational questions for CACS include whether terms which involve intentionality provide an appropriate theoretical framework for the explanation of behavior, and whether intentionality is appropriately attributed to (non-human) animals or to computational models of cognition.

Philosophers often attempt to provide conceptual grounds for the view that intentional states are not appropriately attributed to animals or to computers. Such arguments start with a conception of intentionality and then attempt to show how the given conception applies (or fails to apply) to animals or to computers. For example, some authors have argued that language is a prerequisite for intentionality and that non-human animals thus lack intentionality. Davidson (1975) argues that a creature must have the concept of belief in order to have a belief, and that one cannot have the concept of belief unless one has the capacity for interpreting linguistic utterances. Davidson's position is based on a philosophically sophisticated argument about the nature of interpretation that cannot, for reasons of space, be addressed here. However the upshot is clear: language use provides a criterion for the application of intentional terms. Leahy (1991) uses language in a less sophisticated way to get to the same point. He argues that, lacking a language, animals do not have the kind of behavioral capacities that underpin talk about human beliefs and that it is therefore inappropriate to describe animals as having beliefs. Whether or not language use is as important as these authors believe, it nonetheless provides one criterion for comparative judgments about cognitive abilities.

Other authors have suggested different criteria. Searle (1980) is notorious among computer scientists for arguing, on conceptual grounds, that a program capable of passing the Turing Test need not possess intentionality. His criterion hinges on a notion of "understanding" that is not entirely clear. He thinks that human mental states always involve understanding, and as a result they are intrinsically intentional. Searle argues that, in contrast, execution of program sophisticated enough to pass the Turing test can occur without understanding, so attributions of intentional states to computers in virtue of program execution are, at best, metaphorically derived from the intentionality of those who use them. Searle's view will be discussed in more detail in section 6 below. Dennett (1969) has also argued that the intentionality of computers is derived, not intrinsic. More recently, Dennett (1987) has argued that not even the intentionality of humans is intrinsic—that it is derived from the role of the human body as survival machine for its genes (Dawkins 1976). The argument, roughly, is that we can't say what our mental states are *about* without knowing what they are *for*, in evolutionary terms.

Dennett's view is influenced by Millikan (1984) who bases attributions of intentionality on a biological notion of function. Millikan's view is discussed in section 6.

In this paper, I will argue that cognitive science in general and CACS in particular need not be committed to any particular philosophical view of the nature of intentionality (despite the exhortations of various philosophers to jump on their respective bandwagons). This does not mean that cognitive scientists should ignore philosophical work on intentionality. In fact, to ignore the arguments of someone like Searle could be perilous since funding and publishing decisions are often affected by conceptual arguments such as his. Secondly, I will attempt to show how the various philosophical conceptions of intentionality can provide starting points for comparative, empirical investigations. In doing so, I hope to address criticism such as that by Heyes (1987) who knocks Dennett for suggesting too little in the way of experimentation for cognitive scientists to do. While Heyes may be too harsh on Dennett, who has suggested specific experiments (see Dennett 1983 and Cheney & Seyfarth 1990), the general point is well-taken; philosophical work will be most useful to cognitive science when it can be used to suggest empirical research. Even Searle's work, which seems antithetical to the objectives of many artificial intelligence researchers, can be useful in this way. Finally, it will be seen that the approach to cognitive science which naturally emerges from these considerations is a comparative approach.

2 Intentionality

Most contemporary philosophical work on intentionality is influenced by Chisholm's (1957) discussion of work by Brentano (1924). Brentano, borrowed the notion of intentionality from medieval philosophers, and was interested in what they called the 'intentional inexistence' of mental objects, as opposed to the existence of the objects of physical actions. For example, a person cannot ride a horse unless the horse exists, but a person can desire to ride a unicorn or believe in the existence of unicorns even though there are no unicorns. In this way, the objects of beliefs, desires and certain other mental states are unlike the objects of actions such as kicking, riding, or eating. Beliefs, desires, and the other intentional states can represent or be about things that do not exist. (Intentions, in the ordinary sense of 'purposes', are intentional in the philosophical sense. The fact that one intends to capture a unicorn does not entail the existence of a unicorn.) Brentano's thesis was that *only* psychological phenomena exhibit intentionality and that intentionality is thus a distinguishing characteristic of mentality. His characterization of the intentionality of mental states was 'relation to a content' or 'direction upon an object' (Brentano 1924 as quoted by Chisholm 1957, p. 168). A problem with this characterization is that it does not give a clear test for intentionality.

Chisholm sought to clarify Brentano's thesis by focusing on the language used to describe mental phenomena. He proposed to regard intentionality as a property of *sentences* instead of mental states, and he proposed three logical tests to determine whether or not a given sentence is an intentional sentence. One of these logical properties—failure of substitutivity—will suffice to illustrate the idea. In most sentences, if a given phrase can be substituted for another phrase with the same referent, the truth value of the resulting sentence will be the same as that of the original. For example, given that the name 'Margaret Thatcher' and the description 'the first female British prime minister' have the same referent (i.e. Margaret Thatcher = the first British prime minister) then the sentence 'Margaret Thatcher was deposed' has the same truth value as the sentence 'The first female British prime minister was deposed'. This principle of substitutivity is known as 'substitutivity of identicals'. In some contexts substitutivity appears to fail. For example, the sentence 'John believes that Margaret Thatcher was deposed' need not have the same truth value as 'John believes that the first female British prime minister was deposed'—even though Margaret Thatcher was the first female British prime minister, John might not know it. Chisholm proposed to reformulate Brentano's thesis as the claim that psychological phenomena need to be described using intentional sentences whereas non-psychological phenomena can always be described using non-intentional sentences.

If Chisholm's test worked, it would provide a clear criterion for demarcating psychological phenomena from other phenomena. Unfortunately, substitutivity appears to fail in contexts that are not psychological. For example, although the number of planets = 9, and 9 is necessarily greater than 7, it is not true that the number of planets is necessarily greater than 7—there could have been fewer planets (Quine 1953). In the face of examples like these, failure of substitutivity seems to be an inadequate criterion for psychological intentionality, and Chisholm himself came to this view (Chisholm

1967) although there are philosophers who believe that the apparent counterexamples are not genuine and that the criterion can be saved (Jacquette 1986).

Nowadays, failure of substitutivity is seen as an aspect of the *intensionality* (with an 's') of meaning. (It is easy to confuse the terms 'intentional' and 'intensional' so from here on, where 'intenTion' and 'intenSion' appear in the same or adjacent sentences, I will capitalize the distinguishing letter as shown.) The term 'intenSion' has its major contemporary use by those who attempt to give a formal theory of meaning. Frege (1892) was driven by considerations (including failure of substitutivity) to postulate two components of meaning—sense and reference. Set theory, which Frege helped to develop, provides a powerful apparatus for modeling the referential aspect of meaning. For example, given a domain of individuals, the referent of a proper name, if it has one, is an element in the domain. The referential component of the meaning of a predicate is just the subset of members of the domain to which the predicate applies, called the *extension* of the predicate. In other words, the extension of the predicate 'is green' in a given domain is just the subset of green objects from the domain. On this account, a sentence such as '*a* is green' is true just in case *a* is in the extension of the predicate 'is green'. Extensional accounts of meaning are incomplete. For example, the predicate 'is a member of a species that has evolved kidneys' has exactly the same extension as the predicate 'is a member of a species that has evolved hearts'—the same entities belong to each set. Despite their extensional equivalence, they do not have the same meaning. Since the referential component of meaning cannot account for all semantic phenomena, Frege postulated a second component to meaning, which he called 'sense'. Carnap (1928) is generally credited with introducing the term 'intension' to contrast the second component of meaning with 'extension'.

Quine is a notorious critic of intensions. As he once put it, 'Intensions are creatures of darkness, and I shall rejoice with the reader when they are exorcised...' (Quine 1966, p. 186). The argument for the existence of intensions is theoretical—they appear to be needed to account for certain phenomena surrounding meaning, but there is little agreement about what exactly they are. Whether intenSions can be made philosophically respectable and how they might be relevant to intenTionality remain areas of active philosophical research. The suggestion that there is a connection is tempting—the fact that a person *believes* he is a member of a species that has evolved hearts does not entail that he believes he is a member of a species that has evolved kidneys, despite the extensional identity of the two predicates. The failure of substitutivity in the context of Joe's belief (an inTentional context) might be accounted for by the difference in intenSion of the two predicates, if we only had an adequate account of intenSions.

To non-philosophers, lack of agreement among philosophers about intenTionality and intenSionality may seem to make these notions unsuitable for empirical study. In the rest of this paper I will focus on intenTionality and argue that despite the philosophical controversy, cognitive scientists can make use of the theories that have been proposed.

3 Methodology

Rather than seeing disagreement about the correct account of intentionality as a problem for cognitive science, I see it as an opportunity for developing an empirical account of intentionality. This attitude may seem to present a problem. We are a long way from being able to give an uncontroversial definition of intentionality, but many behavioral scientists believe it is not possible to study a given phenomenon without a rigorous (preferably operational) definition of that phenomenon. That this idea is false should be made obvious by considering early investigation of the chemical nature of elements like gold or carbon. Prior to understanding atomic structure, overt properties such as density, hardness, color, and reactivity, were used to determine whether a given specimen was indeed gold. It would have been premature to define gold in terms of those kinds of properties, since, like carbon, gold could have turned out to occur in more than one form. A precise definition of 'gold' prior to comparative work on numerous putative examples of gold would have begged certain questions about the nature of gold, since *by definition* things that shared the overt properties would have been gold and things that lacked the properties would not have been gold. Rough characterizations in terms of overt properties are not properly regarded as definitions, although they do provide an initial classificatory scheme which is then revised by careful comparative work. The comparative work revised the concept of gold to include ideas about atomic structure, thus legitimizing the concept by fitting into an empirically productive theoretical framework. (See Kripke (1972) for a general account of scientific terms on which these considerations are based.)

The motivation for a comparative approach to the study of intentionality should now be clear. The empirical utility of a notion of intentionality will depend on whether it can be fit into an appropriate theoretical framework. This cannot be decided *a priori* by philosophers any more than philosophers could have decided whether gold, carbon, etc. was a better classification scheme than earth, air, fire, and water. This suggests that cognitive scientists would be ill-advised to look to philosophers for a crisp and empirically rigorous definition of intentionality (even if some philosophers promise to provide it). Philosophical conceptions of intentionality distinguish a certain class of phenomena from others. Given a particular classification scheme based on a particular philosophical conception, further investigation may show whether there is a scientifically useful theoretical basis for including all the phenomena initially characterized in this way. Phenomena initially included may come to be dropped from the categorization scheme, and some phenomena initially omitted may be usefully included. Or the phenomena picked out by the philosophical categories may turn out to be so heterogeneous that no useful theory can be built around them. From this perspective, the variety of philosophical views about intentionality is a good thing insofar as they suggest different bases for comparative studies.

Empirical investigation of a philosophical classification scheme need not commit one to the entire position of that philosopher.[1] For example, the question of whether to treat intentionality as a property of sentences or as a property of minds is less significant from this perspective. Both Brentano and Chisholm provide criteria for distinguishing some phenomena (intentional ones) from others (the non-intentional). The resulting categorization schemes may not precisely overlap but both can provide a starting point for more detailed comparisons of the phenomena. The results of the comparative work may lead to refinements in the notion of intentionality, or to its abandonment, but this cannot be predicted reliably without doing the empirical work. The main point here, though, is that choosing to start from a particular categorization scheme does not commit one to accepting that it is the correct scheme. Indeed, investigation of conflicting categorization schemes might even hasten convergence on a more useful one.

[1] This point is made in detail with respect to cognitive ethology and Millikan's analysis of intentionality, in Bekoff & Allen 1992. Millikan's account is also discussed below.

4 Intentionality and science

Some critics think that there are good philosophical arguments against the possibility of a scientific theory of intentionality. Before addressing some specific criticisms, let us forestall a range of possible arguments. Any valid argument against the possibility of giving a scientific account of intentionality will rely on particular features of a conception of intentionality, so the argument will be, at best, an argument against conceptions that share those features. If there is no guarantee that a scientific theory of intentionality will incorporate those features, then such arguments can show only the unsuitability of a particular conception of intentionality for scientific development, but they cannot constitute an argument against scientific accounts of intentionality in general. Thus it may be possible to respond against some arguments that they rely on too narrow a conception of intentionality. This response strategy, in effect, shifts the target away from the initial criticism and is reasonable up to a point. There are, however, some features of philosophical conceptions of intentionality that may be deemed crucial in the sense that any account of "intentionality" that abandoned those features would be using the same name for something else (such as happened with the appropriation of the ancient Greek term 'atom' by modern physics). If there are central features to the notion of intentionality, and if philosophical arguments against a science of intentionality can be based on those features, then those arguments must be responded to directly.

If there is one feature of intentionality which falls into this category, it is Brentano's notion that psychological states involve content. In English we specify the content of mental states such as beliefs with embedded propositional clauses (as in 'John believes that Mary is coming to dinner'). Dennett (1969), Stich (1983), and Rosenberg (1990) have all objected to basing cognitive science on the notion of mental content on the grounds that in many cases content cannot be specified precisely enough, particularly when considering non-human animals. These authors draw attention to the unsuitability of using a human language to describe the contents of a dog's "beliefs" on the grounds that the human words used will refer to distinctions that humans make but dogs do not. For example, the concept of squirrel has associations for humans that a dog may be insensitive to, that squirrels are phylogenetically close to rats for example. Dennett, Stich, and Rosenberg attempt to use considerations like this to call into question the scientific value of attributions such as ‘Fido believes that a squirrel ran up the tree’ to dogs. But the arguments are really quite weak. Given good comparative data on the discriminative abilities of dogs and humans, there’s no principled reason for thinking that we can’t use human language in some (perhaps) long-winded way to specify the content of a dog’s beliefs.[2]

Comparative approaches are particularly relevant here. Careful comparison of the cognitive mechanisms of dogs and humans will enable us to say how similar their conceptual schemes are to ours, in respect, for example, to associations between various perceptual categories. Without such comparisons, the use of the term ‘squirrel’ in the description of Fido’s belief is anthropomorphic. With such comparisons in hand, it is possible to describe Fido’s beliefs in a way that is not overly anthropomorphic since our descriptions can explicitly cancel connotations that the concept of a squirrel has for normal speakers of English. Once again, the utility of doing this ought not to be decided on philosophical grounds alone, but CACS provides the best approach to deciding the utility of such attributions.

This does not exhaust the challenges to incorporating intentional language into science. One particularly pressing problem is the apparent circularity of attributing beliefs and desires on the basis of actions and then using those belief and desire attributions to predict or explain action. While actions are predictable given knowledge of the contents of beliefs and desires, the contents of beliefs and desires can only be attributed on the basis of actions. If the action explained by a particular set of beliefs and desires is the same action used to justify attributing those beliefs and desires, then there is a circle. But in general our attributions and explanations of behavior are not so tightly linked. If the fact that Joe is carrying his umbrella is our only evidence that he thinks it is going to rain, it is obviously circular to explain his carrying an umbrella by citing his belief that it is going to rain. Usually, however, we have independent evidence that Joe believes it is going to rain, such as his words of warning to his friends, his setting houseplants out in the yard, and so on, and in such circumstances it is not obviously circular to explain his picking up his umbrella by his belief that it will rain. Some critics resort to pointing out that any belief can be used in the explanation of any action, no matter how bizarre the connection, since other beliefs and

[2] The argument against Dennett and Stich is presented in much greater detail in Allen 1992b.

desires can always be attributed to fill the gap. Here's an example from Rosenberg (1988, p. 34): 'Thus, someone might light a cigarette because, say, he believed that the theory of relativity is false. How is this possible? Well, suppose he believed that someone was asking him whether the theory of relativity was true and he also believed that the way to signal dissent in the language of the questioner was to light up and that he wanted to so signal.' Clearly this is possible, but just as clearly we would not take this one action as compelling evidence that the person had the beliefs mentioned.

Here, too, a comparative approach to intentionality is helpful. Ethological and neuroethological work is capable of revealing what features of the environment a given organism's sense organs and nervous system can respond to. Given this kind of knowledge, the range of intentional states that may plausibly be attributed to an organism is constrained. Rosenberg's point that beliefs and desires do not predict anything in isolation is well taken. But this does not justify his pessimistic continuation: 'The number of specific beliefs and desires that lead to actions is so large, and the difficulty of identifying them exactly is so great, that our explanations of action cannot help but be seriously incomplete. ...And our predictions must be equally weak, for they rest on nothing but guesswork about the vast number of specific beliefs and desires that are needed for a precise prediction...' (Rosenberg 1988, p. 34). The range of beliefs that may be attributed to an organism, including a human, is constrained by what we know about its cognitive capacities. In most cases, it is not simply guesswork to discount the idea that someone believes that the appropriate way to signal dissent to a question is to light a cigarette. Such a belief can only be held by an organism with certain perceptual and cognitive abilities, and would only be held as the result of some specific experiences. If we have evidence that either of these conditions is not met then it is not guesswork to withhold the attribution of such a belief.

5 Naturalizing intentionality

I will presuppose that a scientifically useful notion of intentionality must be naturalistic. In other words, it should not require the adoption of any kind of dualism of mental and physical substance. This presupposition provides a tension for attempts to develop a theory of intentionality. On the one hand, there is a tendency to want to assimilate intentionality to physical or biological mechanisms in order to show there is nothing special about it (e.g. Dretske 1981, Millikan 1984, Dennett 1987). On the other hand, we are faced with intuitions about qualitative differences in introspective experiences associated with our mental capacities and the functional capacities of other organisms (e.g., Nagel 1974) and artifacts, such as computers and vending machines (e.g., Searle 1980). Various degrees of obviousness attach to the claims made about such qualitative differences, particularly where other organisms are concerned. Nonetheless, such considerations provide a source of intuitions about the specialness of human cognition and provide a basis for objections to the use of folk psychological terms to describe computers (Searle 1980) and animals (Leahy 1991). CACS has the potential to alleviate the unease arising from the interaction of these intuitions and naturalism, if, against a background of detailed comparison of humans to other organisms and artifacts, neither the similarities nor the differences provide grounds for blanket declarations about the uniqueness, or lack thereof, of human minds. Thus, for example, worries about lack of a language in animals seem less significant when seen as just one dimension along which comparisons can be made.

6 Representations and intentionality

The notion of representation is ubiquitous in the various sciences making up the cognitive sciences (Allen 1992a). This suggests that CACS should exploit this common ground in making comparisons between humans, other organisms, and computers. Comparison of the role of representations in these various systems can ground useful theoretical claims about them all.

The basic notion underlying representation is that of a mapping from the features of one structure to the features of another. Roitblat (1982) develops what he calls a "metatheory" of representation which usefully analyzes the basic notion of a mapping as having four components: a domain (what the representation is used for); content (what is represented); code (the mapping rules); and medium (the physical basis for representation). From a different perspective, Swoyer (1991) gives a useful account of the logical properties of structural representations.

The basic idea of a mapping is widely applicable. In a connectionist network, for example, there may be a mapping between connection strengths and the frequencies with which inputs to the network are correlated. However, not every mapping that exists is of interest to cognitive science. For example, ridges in sand underwater near the shore represent wave patterns because there is a mapping from ridge orientation to average wave direction. But there is no need to think of the ocean floor as a cognitive device, or displaying intentionality. The same point can be put in terms of information. In the sense of 'information' defined by Shannon & Weaver (1949), the ridges in the sandy ocean floor are a source of information about the waves. Cognition is often characterized as information processing, but, clearly, not all information bearing structures are cognitive. Philosophers are interested in developing notions of information that are richer than the Shannon & Weaver notion and that can be used to suggest empirical work for comparative cognitive science (Dretske 1981; Allen & Hauser 1993).

Although the tasks of developing cognitively interesting notions of representation and information are related, to pursue both topics would take us too far afield here (but see Allen & Hauser 1993). The remainder of this section focuses on the question of how to extend the basic notion of representation in a way that can be applied to questions about intentionality.

6.1 Intentionality and control

The first suggestion is that a representation is intentional if it is used by a system to control behavior. The idea that *behavioral control* is a key feature of cognitive representation is implicit in Roitblat's (1982) characterization of the domain of a representational system as the "class of situations or tasks in which it is used and to which it applies" (p. 353). Roitblat does not discuss intentionality explicitly, but the notion of representation he develops is an intentional notion in the sense that representations are *about* features of the represented world. (Reader exercise: Are they intenSional?)

The idea that the intentionality of a representation derives from what it does in that system is "functionalist" in the sense of Cummins (1989, ch. 9). In a functionalist account of representation, the functional role of a given representation may be specified either causally (i.e., by the role the representations play in the causal network between environmental stimuli and behavior), or computationally (i.e., but the role such representations play in a Turing-machine characterization of the system). But, as Cummins points out, it is convenient to use the term "functionalism" generically without specifying in which way functional roles are to be attributed.

Functionalism underlies a common view that the mechanisms underlying human intentionality are just very complicated versions of those found in common artifacts. Thus, for example, McCarthy (1980) argues that it is appropriate to attribute beliefs to thermostats and Dretske (1980) argues that it is appropriate to attribute intentionality to similar devices. Thermostats contain mechanical parts (e.g., metal coils) whose properties (e.g., coil tension) map the temperature of the environment and directly control the heating and cooling devices. Coil tension maps onto ambient temperature, thus representing the temperature. Other features of the thermostat may also represent the temperature (such as the degree of expansion of the plastic housing) but whereas coil tension features in the causal story about furnace switching, expansion of the housing does not. Thus, according to this suggestion, it is the role of a representation in producing behavior that constitutes its intentionality.

This approach is obviously compatible with the desire to naturalize intentionality. However, the approach also seems to run afoul of intuitions against attributing mental properties to pieces of metal and plastic commonly found on walls. Thus, for example Searle (1980) vehemently disagrees that intentional descriptions of thermostats are comparable to belief attributions in humans. My purpose here does not include adjudicating this dispute. Models of human cognitive performance which focus on the representational requirements of brain mechanisms provide an obvious domain of comparison with similar models of animal competence on with artificial intelligence models. CACS can make use of such comparisons without prior commitment to the correctness of one philosophical position over another.

6.2 Biofunctional intentionality

The second suggestion (Millikan 1984) is that a representation is intentional if similar mappings have historically played a role in the survival of ancestral systems. Millikan's account of intentionality is perhaps the most detailed attempt to put intentionality into a biological framework, and it is beginning to attract attention from ethologists (e.g., Beer 1991, Bekoff & Allen 1992). On Millikan's view, intentionality is a biological property derived from the biological functions of those things that possess it. She uses bee dances as examples of what she calls 'intentional icons'. According to her account, bee dances are about the location of nectar because the adaptive value of bee dances for ancestors of current bees is explained by a representational mapping between features of the dance and the location of nectar.

Millikan's account is historical in the sense that a thing's intentionality depends not on present characteristics but on it being the product of a selective process which allows us to say what the thing is for. This leads her to claim that an exact duplicate of a human being produced by a random process (e.g. an extremely unlikely quantum accident) while it might be conscious or have other mental states would not have any intentional states such as beliefs, desires, etc. since, initially at least, representational properties of these states have nothing to do with their existence (Millikan 1984, p. 93). It is important to realize that the sense of 'function' invoked by Millikan is very different from that inherent in the functionalism of the previous suggestion. In most functionalist accounts, it does not matter what the history of a system is—all that counts is its current Turing-machine description. If it has certain functional capabilities according to this description, then it has intentionality. On Millikan's account, however, it is the historical role of similar states in adapting other systems to their environments that determine function and thus intentionality. This is why she thinks two devices can be identical in all physical respects yet differ in that one has intentionality and the other does not.

Applying her account of intentionality, even trees can exhibit it. For example, in work on African acacia trees it has been found that individual trees will increase tannin production in their leaves in response to predation by kudu antelope to a level that can kill the antelope. In addition to this primary response, acacias release ethylene into the air causing downwind acacias up to 50 yards away to step up their tannin production within 5 to 10 minutes. The adaptive significance of ethylene release is obvious and has led at least one science journalist to call it an 'alarm system' (Hughes 1990).[3] For acacias, ethylene has historically played the role of adapting downwind trees to the environmental condition of predation by herbivores. Hence, ethylene released by the trees is an intentional icon on Millikan's account.

[3] Hughes reports that this phenomenon was discovered by Wouter Van Hoven after the sudden death of approximately 3000 kudu. She also reports his observation that giraffe feeding on acacias only graze on about 1 tree in 10 and avoid downwind trees.

These examples point out the distance between Millikan's account and Brentano's view that intentionality is the distinguishing mark of the mental. She admits that an entity might be conscious while lacking intentionality. On the other hand, on the plausible assumption that acacia trees don't have minds, her account allows intentionality to occur in conjunction with non-psychological phenomena. I am not arguing here whether it is a good thing or a bad thing that Millikan's account of intentionality diverges so much from Brentano's or whether CACS should adopt one view rather than another. Instead, I wish again to emphasize my point that divergence between philosophers on the subject of intentionality need not be considered a hindrance to cognitive scientists. Whether or not one thinks that Millikan's account of intentionality is correct, its focus on historical antecedents suggests a line of comparison that other accounts would not.[4] On Millikan's view, intentional content is not specified by looking inside the system in question, but by looking at how the system is adapted to its environment. Insofar as this perspective can be applied to organisms, robots, or other entities, it provides a basis for comparison.

Is this the right basis for a comparative approach to intentionality? By now it should be clear that I will not answer this question, on the grounds that such questions should be addressed empirically. Conceptual arguments have their place, but the real test of an approach is in its application. The question also contains a problematic presupposition, namely, that there is a single correct basis for comparing the cognitive abilities of diverse organisms and machines. This presupposition may be unwarranted since the purposes of different scientists, even within the various cognitive sciences, are divergent. Given my exact physical constitution at this moment, whether I came together as a quantum accident two seconds ago or whether I came to be typing this paper by a more normal route does not bear on the probability of my ending this sentence with a period. A theory that focuses on predictive aims such as this one might have little use for categorizing things in terms of their evolutionary history. For other purposes, comparison on the basis of this history might well turn out to be significant. Different sciences with different purposes can coexist. It might even turn out to be useful for some purposes to categorize the behavior of trees with the behavior of humans,

[4] Bekoff & Allen (1992) discuss the application of Millikan's theory to canid play behavior.

animals, and computers.

6.3 Intentionality and detection of error

The third suggestion is that a representation is intentional if the system in which it plays a role is capable of discriminating content from reality. On this test, thermostat coils fail to be intentional representations of temperature since the thermostat cannot distinguish circumstances where tension is applied manually to the coil (and thus does not correspond to the ambient temperature) from those where the coil tension properly represents ambient temperature. In comparison, humans and some organisms are capable of treating their own perceptual states as misleading. I would suggest that how this capability is realized is of major interest for comparative cognitive science.

This suggestion, like that in section 6.1 above, is functionalist because it looks to the role representations play for the present capabilities of the system in question. However, it differs from the suggestion in section 6.1 by requiring a more sophisticated role for the representations to play. In doing so it avoids some of the pitfalls of Searle's challenge to the possibility of realizing artificial intentionality via a computer program. He has a conception of intentionality as a property of mental states directly, but precisely what property is not entirely clear from his account. According to Searle, intentionality requires 'some awareness of the causal relation between the symbol and the referent' (Searle 1980, p. 454) but it is unclear what this awareness amounts to. He also suggests that intentionality inherently depends on the specific biochemical properties of brains (Searle 1980, p. 424). Searle gives no indication which biochemical properties of brains are necessary, nor does he give any idea why biochemistry might be important to intentionality—he makes a vague analogy with digestion, but with digestion it is clear why biochemical properties are important. Furthermore, Searle rejects all forms of the Turing Test as evidence of intentionality. Thus, absolutely nothing an A.I. system could do would convince him that the system possessed intentionality. This has the unfortunate consequence of removing the question of the intentionality of A.I. systems from the empirical realm.

This idea that intentionality requires not just a causal or computational connection between symbol and referent, but awareness of that connection is compatible, however, with the suggestion at the top this section. We can use a comparative approach to cognitive science to demystify the notion of awareness here, by considering what it is about humans that makes us attribute such awareness to each other and then trying to build similar capabilities into computers and trying to find those capabilities in other organisms. One of the things that convinces us of this awareness, I believe, is that humans sometimes are skeptical of the veracity of their immediate perceptual stimuli. What other organisms are capable of similar skepticism, and over what range of their potential stimulations they are capable of such skepticism is a prime question for empirical research. An anecdote about vervet monkeys can be used to illustrate the point. Subordinate male vervets isolated with an infant behave less aggressively toward the infant when they can see they are being watched by the infant's mother who is isolated behind a clear piece of glass, than when the same female is hidden behind a piece of sheet metal or behind a one-way mirror. When the females are subsequently released to interact with the males, female aggression toward the males is correlated with what they have seen through the glass or the one-way mirror (Keddy-Hector et al. 1989). After only a few experiences, males placed in a cage with a one-way mirror spent more time inspecting the mirror than interacting with the infant (Keddy-Hector, reported in conversation). This anecdotal evidence is suggestive since it appears to indicate that vervets can distinguish the appearance of not being watched from the actuality of not being watched. This kind of observation makes it plausible to include beliefs about whether they are being watched or not as part of the explanation of the behavior of the males toward females. The possibility of discovering this kind of capability in different species further illustrates how it is possible to address Rosenberg's worry about the guesswork involved in attributing the background beliefs needed to explain an action.

The suggestion of this section also makes the variety of sensory modalities found in cognitively sophisticated organisms particularly interesting. Humans can use one sensory modality, e.g. touch, as a test of their perceptions in another modality, e.g. vision, and also use perceptions in a given modality to test earlier perceptions from that same modality. Apparently, the human brain produces a model of the world and then tests its perceptions against that model. Investigating the computational requirements for such a capability is a challenging task for A.I. that, as far as I know, has not been attempted. The degree to which organisms can question its immediate perceptions is a topic whose empirical investigation is relevant to the conception of intentionality considered in this section.

7 Summary

No one of the accounts of intentionality that I present here is suggested as a full analysis of intentionality. Instead, each indicates avenues of research, both for observation of animals and for development of computational models. Cognitive scientists should not look to philosophers for a prepackaged concept of intentionality. Rather, the different philosophical theories which are available can be used to motivate various empirical lines of enquiry. By using a particular philosophical theory in this way one is not committed to all of the philosophical baggage that comes with that theory. In the end, it may turn out that intentionality is not a useful concept for cognitive science, or it may turn out that it is a useful concept. In either case, the decision should not be made on conceptual grounds alone, such as lack of language, lack of a particular biochemical makeup, or lack of an appropriate history, despite what some philosophers may think. Furthermore, different conceptions of intentionality may be useful for different scientific goals. Only a fully empirical comparative approach to cognitive science is adequate to dealing with the topic of intentionality.

Acknowledgements

I would like to thank the organizers of CACS 92, particularly Jean Arcady-Meyer and Herb Roitblat, for their invitation and efforts to make the summer school a success. I am grateful to the French CNRS for its financial support of CACS 92. Marc Bekoff and I have spent many hours discussing these and other issues. He, Dale Jamieson, Ruth Millikan, and Herb Roitblat all provided helpful comments on the manuscript.

References

[1] Allen, C. (1992a) "Mental content and evolutionary explanation." *Biology and Philosophy* **7**: 112.

[2] Allen, C. (1992b) "Mental content." *British Journal for the Philosophy of Science* **43**: 537-553.

[3] Allen, C. & Hauser, M. (1993) "Communication and cognition: is information the connection?" *PSA 1992, vol. 2*: 81-91.

[4] Beer, C. G. (1991) "From folk psychology to cognitive ethology." In Ristau ed. (1991): 19-34.

[5] Bekoff, M. & Allen, C. (1992) "Intentional icons: towards an evolutionary cognitive ethology." *Ethology* **91**: 1-16.

[6] Brentano, F. (1924) *Psychologic vom empirischen Standpunkte.* Leipzig: F. Meiner.

[7] Carnap, R. (1928/1967) "The logical structure of the world: pseudoproblems in philosophy." Translated by R. A. George. Berkeley, CA: University of California Press.

[8] Cheney, D. L. & Seyfarth, R. M. (1990) *How monkeys see the world: inside the mind of another species.* Chicago: University of Chicago Press.

[9] Chisholm, R. M. (1957) *Perceiving: a philosophical study.* Ithaca, NY: Cornell University Press.

[10] Chisholm, R. M. (1967) "Intentionality." In Edwards, P. (ed.) *The Encyclopedia of Philosophy*, vol IV: 203. New York: Macmillan and The Free Press.

[11] Churchland, P. S. (1986) *Neurophilosophy: toward a unified science of the mind/brain.* Cambridge, MA: MIT Press.

[12] Cummins, R. (1989) *Meaning and mental representation.* Cambridge, MA: MIT Press.

[13] Davidson, D. (1975) "Thought and talk." In Guttenplan, S. (ed.) *Mind and language.* Oxford: Oxford University Press.

[14] Dawkins, R. (1976) *The selfish gene.* Oxford: Oxford University Press.

[15] Dennett, D. C. (1969) *Content and consciousness.* London: Routledge & Kegan Paul.

[16] Dennett, D. C. (1983) "Intentional systems in cognitive ethology: The 'Panglossian paradigm' defended." *Behavioral and Brain Sciences* **6**: 343-345.

[17] Dennett, D. C. (1987) *The intentional stance.* Cambridge, MA: MIT Press.

[18] Dretske, F. I. (1980) "The intentionality of cognitive states." In P. A. French, T. E. Uehling Jr. & H. K. Wettstein (eds.) *Midwest Studies in Philosophy*, vol V: 281-294. Minneapolis: University of Minnesota Press.

[19] Dretske, F.I. (1981) *Knowledge and the flow of information.* Cambridge, MA: MIT Press.

[20] Frege, G. (1892/1948) "Über Sinn und Bedeutung." Translated as "On Sense and Reference" by M. Black. *Philosophical Review* **57** (1948): 209-230.

[21] Heyes, C. (1987) "Contrasting approaches to the legitimation of intentional language within comparative psychology." *Behaviorism* **15**: 41-50.

[22] Hughes, S. (1990) "Antelope activate the acacia's alarm system." *New Scientist* **127**: 19 (Sep 29 1990).

[23] Jacquette, D. (1986) "Intentionality and intensionality: quotation contexts and the modal wedge." *The Monist* **69**: 598-608.

[24] Keddy-Hector, A. C., Seyfarth, R. M. & Raleigh, M. J. (1989) "Male parental care, female choice and the effect of an audience in vervet monkeys." *Animal Behavior* **38**: 262-271.

[25] Kripke, S. A. (1972) *Naming and necessity.* Cambridge, MA: Harvard University Press.

[26] Leahy, M. P. T. (1991) *Against liberation: putting animals in perspective.* London: Routledge.

[27] McCarthy, J. (1980) “Beliefs, machines, and theories.” *Behavioral and Brain Sciences* **3**: 435.

[28] Millikan, R. G. (1984) *Language, thought, and other biological categories.* Cambridge, MA: MIT Press.

[29] Nagel, T. (1974) “What is it like to be a bat?” *Philosophical Review* **83**: 435-50.

[30] Quine, W. V. O. (1953) *From a logical point of view.* Cambridge, MA: Harvard University Press.

[31] Quine, W. V. O. (1966) “Quantifiers and propositional attitudes.” In Quine, W. V. O., *The ways of paradox and other essays*: 185-196. Cambridge, MA: Harvard University Press.

[32] Roitblat, H. L. (1982) “The meaning of representation in animal memory.” *Behavioral and Brain Sciences* **5**: 353-406.

[33] Rosenberg, A. (1988) *Philosophy of social science.* Boulder, CO: Westview Press.

[34] Rosenberg, A. (1990) “Is there an evolutionary biology of play?” In Bekoff & Jamieson, eds. (1990) vol. II: 180-196.

[35] Searle, J. R. (1980) “Minds, brains and programs.” *Behavioral and Brain Sciences* **3**: 417-424.

[36] Shannon, C. E., & Weaver, W. (1949) *The mathematical theory of communication.* Urbana, Illinois: University of Illinois Press.

[37] Stich, S. (1983) *From folk psychology to cognitive science: the case against belief.* Cambridge, MA: MIT Press.

[38] Swoyer, C. (1991) “Structural representation and surrogative reasoning.” *Synthese* **87**: 449-508.