Mining the Past to Construct the Future: Memory and Belief as Forms of Knowledge

1

Chris Westbury
Daniel C. Dennett

> The analogy between memory and a repository, and between remembering and retaining, is obvious and is to be found in all languages; it being natural to express the operations of the mind by images taken from things material. But in philosophy we ought to draw aside the veil of imagery, and to view them naked.
>
> *Thomas Reid,*
> Essays on the Intellectual Powers of Man *(1815)*

Jacques Monod (1974) observed that "ever since its birth in the Ionian islands almost three thousand years ago, Western philosophy has been divided between two seemingly opposed attitudes. According to one of them the true and ultimate reality of the universe can reside only in perfectly immutable forms, unvarying by essence. According to the other, the only real truth resides in flux and evolution" (p. 98). Three thousand years of argument has so far failed to find a clear resolution to this ancient opposition. There are still two committed camps of "Neats" and "Scruffies," who fall on either side of the fundamental debate identified by Monod. Some of the most exciting points of intersection between the interests of philosophers and scientists today are those at which the apparent incompatibility between the two camps demands to be resolved—those points at which it becomes clear that the stable, neat objects of scientific inquiry are attaining their objective status by managing to separate themselves (often using scruffy means) from an underlying scruffy flux of dynamic phenomena. When we try to understand such objects, questions of science merge unavoidably with questions of epistemology. The questions that interest us about memory and belief exist at this intersection. We will briefly consider each of these phenomena in turn. In doing so, we will try to point out some conceptual confusions that spring from the way we use their names in informal discourse, and emphasize a strong underlying similarity in the ways that memory and belief relate to knowledge.

Memory

Every event in the world has effects, and the chain of effects that spreads from any event continues essentially forever; but only some events leave long-lasting traces. We single out the best cases of this for special notice: footprints, scars, and various sorts of records. Of all the dinosaur footprints that ever pressed into mud, only a tiny fraction are discernible today; of all the clay tablets ever impressed with hieroglyphics, only a select few survive—but like the dinosaur footprints, they permit us to read the past in a way that the other long-lived effects of the same causes do not. Fifty light-years from Earth, a sphere of 1947 Jack Benny broadcasts is expanding into the galaxy, almost certainly unreconstructible by any technology. If those programs were not recorded here on Earth, they would be gone forever. Events that leave no salient long-term traces can be called inert historical facts; they happened, but the difference they made no longer makes a discernible difference. There is no fixed best way to count facts, but by almost any usable method we would have to say that most historical facts are inert. One of the following is a fact: (a) some of the gold in Dennett's teeth once belonged to Julius Caesar; (b) none of the gold in Dennett's teeth ever belonged to Julius Caesar. Although one of these statements is true, it is almost certainly beyond all powers of investigation to determine which.

The past consists of all historical facts, inert or recoverable. The whole point of brains, of nervous systems and sense organs, is to produce future, to permit organisms to develop, in real time, anticipations of what is likely to happen next, the better to deal with it. The only way—the only nonmagical way—organisms can do this is by prospecting and then mining the present for the precious ore of historical facts, the raw materials that are then refined into anticipations of the future. As Norbert Wiener (1948) pointed out long ago, the fundamental method is trajectory sampling or tracking: gathering data about the pattern of change in something of interest, then extrapolating the curve into the future. Whether an organism tracks temperature, or salinity, or the sun, or a prey, or a mate, or the Dow-Jones industrial average, the fundamental problem is the same: preventing selected historical facts from going inert, at least long enough to extract their portent for the future.

In his recent book *On the Origin of Objects,* Brian Smith (1996) avails himself of a similar idea of tracking to develop a metaphysical thesis about how the world came to split itself into the experiencer and the experienced, into organism and object. In discussing memory, he writes, "In order to make a memory more durable than a shadow, a . . . subject must first allow or arrange for an appropriate impression to be formed by the event in question, but must then take responsibility for storing it *in such a way as to ensure that it will continue to be effective* after the event it records has dissipated" (p. 221, emphasis added; see also Plotkin, 1994, pp. 149–152). Only organisms that can retrieve stored information in order to increase the likelihood of achieving some adaptive end will gain any advantage from memory. Memory must be rooted in use.

What is the simplest kind of use to which a rudimentary protomemory might be put? Consider the visual perception of motion or change. An organism's visual system could not tell the difference between something moving from left to right and something moving from right to left unless it had *some* way of getting the data from two moments of time into a single comparison process. Similarly, in order to tell that it is getting warmer, not colder, an organism has to have some way of "remembering" the temperature in the recent past. We put "remembering" in scare-quotes, because although this is the most fundamental form of memory, the basis for all others, it need not be much like conscious recollection. Protozoan memory, or the memory in a tree that reminds it to start pushing out buds in the spring, is not like recalling your first-grade teacher. Yet in one fundamental way, such functional use of stored information *is* like memory—that is why we call it memory, however scare-quoted. Memory in the fundamental sense is the ability to store useful information and to retrieve it in precisely those circumstances and that form which allow it to be useful.

Computer memory is memory in this fundamental sense—and only in this fundamental sense, obviously. The process of changing the electromagnetic properties of tiny areas on chips or disks may create many long-lasting local effects, but only those count as being stored in memory that can later be retrieved by the hardware in the ways intended when they were laid down. Any other changes wrought, however salient, long-lasting, or effective in altering the system's behavior, count as blemishes or scars, not memories. Scars are not memories. A dog whose body shows scars inflicted by some encounter but who exhibits no heightened caution, no discrimination of the harbingers of further scars—in short, who has learned nothing from the encounter—has no useful memory of the encounter, even if we others can read a lot about it in our examination of the scars. *We* can extract information from the traces left by the event; the dog, apparently, cannot.

Where do we draw the line? If one of the effects of the earlier encounter is that the dog now limps, and if this limp usefully avoids pain and further injury to the limb, does this count as memory? We need not answer each such question, but it is important to note that a long-lasting and salient trace is not enough, and that what more must be added—utility—comes in different grades and amounts. At one extreme, we find the paradigmatic cases of memory: conscious, reflective beliefs about the past, made manifest in episodes of articulate recollection: "Share your memories of the war with us, Grandpa." "I'd be delighted to: On April 22, 1943—I can see it as if it were this morning—I woke to the sound of mortar fire. . . . " At the other extreme, we find a slight tendency to crouch when a passing car backfires—all that remains of a past, once-useful response.

This twofold requirement—a trace and its utility—raises a quandary that has bedeviled philosophers and psychologists for several millennia. In its starkest form, it was articulated in Plato's *Meno.* In that early dialogue, the wealthy Meno challenges Socrates with an apparently paradoxical question: How is that we can ever discover anything new, given that we must either know what we are looking for (and therefore have no need to look for it) or not know what we are looking for (in which case we will have no way of recognizing it when we find it)? Socrates' answer is the doctrine that all knowledge is *anamnesis:* the recollection of innate knowledge, obtained from previous lives. It is hardly satisfactory. How, Meno might well have gone on to ask, do we tell veridical recollection from sheer fantasy? Reminiscence is problematical in the same way as knowledge: to recall something, we must already know what it is we are trying to recall (and therefore have no reason to recall it) or else not know it (and so have no ability to recall it).

In the *Theaetetus* Plato introduces two metaphors for memory, comparing it to an image inscribed on a wax tablet, which may be more or less smooth, muddy, and soft, and to a bird in an aviary:

> SOCRATES: Now consider whether knowledge is a thing you can possess in that way without having it about you, like a man who has caught some wild birds—pigeons or what not—and keeps them in an aviary for them at home. In a sense, of course, we might say that he "has" them all the time inasmuch as he possesses them, mightn't we?
>
> THEAETETUS: Yes.
>
> SOCRATES: But in another sense he "has" none of them, though he has got control of them, now that he has made them captive in an enclosure of his own; he can take and have hold of them whenever he likes by catching any bird he chooses, and let them go again; and it is open to him to do that as often as he pleases.

Neither metaphor provides a satisfactory answer to the question raised by Meno. If memory is an image carved in a wax tablet, we must somehow know where to look on the wax tablet, or at which wax tablet to look. If it is a bird in an aviary, we must not only be able to call it, but know which one to call and know that it will come when we call. Hotspur retorts to Glendower when he claims (in Shakespeare's *1 Henry IV*) that he "can call spirits from the vasty deep": "Why, so can I, or so can any man; But will they come when you do call for them?" Meno's problem of needing to know what you need to remember in order to remember it still presents a problem for modern theories of memory.

Plato's student Aristotle devoted a short treatise to the problem of memory (Aristotle, c. 350 B.C./1941), in which he proposed that memory was a picture *(eikon)* of the past thing remembered. This picture was causally related to a past object of perception, which was imprinted into a sense organ that was capable of perceiving it. Aristotle's proposal that the object of memory is an accurate picture of a perception, which can simply be consulted as if it were an object of perception, largely reduces the problem of memory to the problem of perception. In remembering *x*, we know that it is *x* for the same reason (whatever that might be!) that we were able to recognize *x* when we first sensed it, because a memory is simply another "viewing" of that same sensory impression. This viewpoint defines what was to become the standard representative theory of memory. Memory came to be seen as what John Locke (1700) called a "Store-house of our Ideas" (p. 150).[1]

There are numerous problems with the representative theory of memory, most of them echoes of Meno's question. A major difficulty is explaining how we can distinguish between imagination and memory. Aristotle's solution was to introduce the capacity of recognizing elapsed time, which allows one to connect the present memory image to an earlier act of perception. His conception of how this might work rested on a visual analogy, that of seeing objects at different distances (Aristotle, c. 350 B.C./1941, p. 615). Earlier memories have receded farther into the distance, and so are smaller and less distinct, than more recent memories.

David Hume (1739) suggested a way of differentiating memories from images that was similar to Aristotle's suggestion. According to Hume, what differentiated memory ideas from imaginative ideas was not a time stamp but the fact that the ideas of imagination are "fainter and more obscure" than the ideas of memory, which strike one with "superior force and vivacity" (p. 85). Thomas Reid (1815) was critical of Hume's view of memory, arguing that if Hume believed (as he claimed to) that ideas had only a contingent, rather than a necessary, relationship, then there could be no deductive argument that depended on the recognition of a relationship between one idea and another (that is, between a sense idea and a memory idea). Hume had himself recognized the difficulty and tried, in an appendix to his book, to replace the idea of "vivacity" with the idea of "apprehending the idea more strongly or taking a firmer hold of it" (Hume, 1739, p. 624). On this attempt, Reid commented dryly, "There is nothing more meritorious in a philosopher than to retract an error upon conviction, but in this instance I humbly apprehend Mr. Hume claims that merit upon too slight a ground" (Reid, 1815, p. 380). Reid's typically commonsensical solution to the problem begged the question by simply accepting that memory, like sense data, was a form of immediate and noninferential knowledge. (How do I tell a memory from a fantasy, Meno? I just do.)

John Stuart Mill (1869) proposed another mechanism for recognizing a memory as a memory. He suggested that in accessing a memory we run very quickly over all sense impressions since the remembered event, and use our understanding of the total number of events to locate the idea in its temporal location.

> In . . . recollection there is, first of all, the ideas or simple conceptions of the object and acts; and along with those ideas, and so closely combined as not to be separable, the idea of my formerly having had those same ideas. And this idea of my formerly having had those ideas is a very complicated idea, including the idea of myself at the present moment remembering, and that of myself of the past moment conceiving; and the whole series of the states of consciousness, which intervened between myself remembering, and myself conceiving. (pp. 330–331)

Many more recent philosophers have followed Thomas Reid in choosing to leave the difficult idea of "pastness" as an unanalyzed primitive. For example, William James wrote simply that memories were referred back in time by "a general feeling of the past direction of time" (James, 1890, p. 650). Similarly, Bertrand Russell (who ultimately remained skeptical of any necessary connection between memory and belief) suggested that a feeling of pastness and a feeling of familiarity gave rise to a feeling of belief, and that all three were essential constituents of memory (1921, pp. 161–163). (For further discussion of how philosophers subscribing to the representative theory have tried to explain the differentiation of memories from imagination, see Holland, 1954.)

Ironically, it is Plato's theory, rejected by Aristotle more than two thousand years ago, which is perhaps closest in spirit to modern theories of memory. Plato's theory of anamnesis is not to be scoffed at, even if we have to reinterpret his central metaphor in order to make any use of the idea. The main point he was making has been rediscovered again and again: there can be no learning—and so no knowledge—from the base of a tabula rasa. It follows that a memory cannot just be an isolated atomic fact, complete unto itself. James (1890) made this point explicitly, noting that

> what we began calling the "image," or "copy," of the fact in the mind, is really not there at all in that simple shape, as a separate idea. Or at least, if it be there as a separate idea, no memory will go with it. What memory goes with is, on the contrary, a very complex representation, that of the fact to be recalled *plus* its associates, the whole forming one "object" . . . and demanding probably a vastly more intricate brain-process than that on which any simple sensorial image depends. (p. 651; emphasis in original)

F. C. Bartlett (1932) would later reemphasize this same point in his criticisms of Ebbinghaus' early attempts to study memory with "meaningless" stimuli. He too recognized that for a memory to be *recognized,* there has to be *independent knowledge* against which it can be compared—how else? Even Skinnerians appreciated the need for an innate basis for distinguishing reinforcers, positive and negative. It was a failure to deal adequately with this necessity for independent knowledge that caused so many problems for the representationalists—they could find nothing to which their memory image might be compared, in order to recognize it *as* a memory. To compare it to another memory image would hardly do, since that would lead to an infinite regress.

Meno's question about how we can recognize our memories has been recast in its original form, as a question about how to recognize our own knowledge. It is a mistake, as everyone acknowledges, to assume that interpretation of recalled knowledge happens only at the preloading, perceptual stage—to assume that once knowledge "enters memory" (through the front door), it is happily stored away, to be "retrieved" intact at later times of "recollection." Though everyone recognizes this as a bad view, it still haunts discussions subliminally.

The importance of interpretation in accessing memory has of course been known for a long time. Starting with Bartlett's (1932) famous experiments examining the recollection of stories, which led him to conclude that organisms have an "effort after meaning," it has been repeatedly demonstrated that our ability to recall is inextricably linked to our assumptions about how the world is, and is subject to "top-down" schematization dictated by those assumptions (see Goldman, 1986, for a review and analysis of empirical evidence for this claim). What we recall is not what we actually experienced, but rather a reconstruction of what we experienced that is consistent with our current goals and our knowledge of the world. As Bartlett put it: "Remembering is not the re-excitation of innumerable fixed, lifeless, and fragmentary forms. It is an imaginative reconstruction, or construction, built out of the relation of our attitude towards a whole active mass of organized past reactions or experience, and to a little outstanding detail which commonly appears in image or in language form" (1932, p. 213, cited in Zechmeister and Nyberg, 1982, p. 301).

The assessment of what constitutes an acceptable reconstruction of the past must be dynamically computed by an organism under the constraints imposed by its built-in biological biases and the history of the interaction of those biases with the environment in which the organism has lived. The apparently stable objects of memory—the representations of the things being recalled—are not retrieved from some Store-house of Ideas where they have been waiting intact, but rather are constructed on the fly by a computational process. As H. R. Maturana (1970) wrote, "Memory as an allusion to a representation in the learning animal of its past experiences is also a description by the observer of his ordered interactions with the observed animal [which may be the observer himself]; memory as a storage of representations of the environment to be used on different occasions in recall does not exist as a neurophysiological function" (p. 37). What we call recollection can never be more than the most plausible story we come up with (or, perhaps, only a story which is plausible enough) within the context of the constraints imposed by biology and history.

Belief

The word "belief" is difficult to define, in part because it is used in different ways in different contexts. In his closing statements at the conference from which the chapters in this volume have been drawn, Antonio Damasio pointed out that a person who talked about his beliefs about everyday things—about the shape or purpose of doorknobs, pencils, telephones, and irons—would be taken for either a comic or a lunatic (or, as one of us added, a philosopher!). Damasio was trying to capture a commonly held intuition: that beliefs cannot be about just anything at all; they must be about important or notably uncertain things. In ordinary usage, the term "belief" is reserved for referring only to linguistically encoded convictions and doctrines, not to the mass of unexpressed and widely held background knowledge that we all implicitly use to navigate through our world.

We would not normally be willing to say that a person who makes himself a cup of coffee is thereby expressing his beliefs about the physical characteristics of electrical outlets, coffee machines, ceramic containers, hot liquids, and organic compounds such as coffee beans, sugar, and milk, or even that he is expressing a belief about the desirability of coffee. He is simply making a cup of coffee. However, within the cognitive science and philosophy of mind communities, the word "belief" is often used much more generally, to refer to any implicit or explicit information that guides an agent's voluntary actions. One benefit of this way of thinking is that it allows us to speak of the beliefs of nonhuman agents (see Dennett, 1987a, 1995, 1996). By widening the meaning of the word, we lose certainty about exactly when it applies. Does the duckling who follows Konrad Lorenz truly believe that the large bearded man is its mother? Does the ant following a pheromone trail or the amoeba swimming up a chemical gradient really believe that following those signals will lead to a food supply? When we see a cow in a field, do we believe simultaneously that our eyes are not fooling us, that an object that looks like a cow is a cow, that our memory for names of animals is functioning properly, that we are not dreaming, and so on for the thousands of other propositions that must all be true in order for us to believe that the cow in the field is indeed precisely that? Are we to accept that any cognitive state must encode all the implicit beliefs which must be held in order for that state to be interpreted properly?

We will consider two possible recent approaches to these kinds of questions: the "language of thought" approach and the "intentional stance" approach.

The Language of Thought

The tradition in cognitive science and philosophy of mind that we characterize as the language of thought tradition shares some philosophical roots with the representative theory of memory. This language of thought tradition has considered beliefs to be individual data structures to be found somewhere in the brain (in the Golden Age of Neurocryptology, presumably). To those who hold this strange view, this has seemed to follow from the fact that beliefs are, in philosophical parlance, "propositional attitudes," instances of the formula

$$x \text{ believes that } p,$$

where p is replaced by some sentence expressing the proposition believed. But we need not, and indeed should not, jump so blithely to the conclusion that the beliefs about which we want to have theories in cognitive science are anything like that. To see why not, consider the following experiment, with yourself as subject:[2]

> Here is a joke. See if you get it. (Newfies are people from Newfoundland; they are the Poles of Canada—or the Irish of Canada, if you're British.)
>
> A man went to visit his friend the Newfie and found him with both ears bandaged. "What happened?" he asked, and the Newfie replied, "I was ironing my shirt, you know, and the telephone rang." "That explains one ear, but what about the other?" "Well, you know, I had to call the doctor!"

The experiment yields a positive result if you get the joke. Most people do, but not all. If we were to pause, in the fashion of Eugene Charniak, whose story-understanding AI (artificial intelligence) program (Charniak, 1974) first explored this phenomenon, and ask what one has to believe in order to get the joke, we would generate a very long list of different propositions. We would need to include propositions about the shape of an iron and the shape of a telephone; about the supposed fact that when people are stupid they often cannot simultaneously do different things with the left hand and the right hand; about the fact that the heft of a telephone receiver and an iron are approximately the same; about the fact that when telephones ring, people generally answer them; and many, many more.

What makes the brief narrative a joke and not just a boring story is that it is radically enthymematic; it leaves out a lot of facts and counts on the listener's filling them in, which the listener is able to do only if she believes all those propositions. Now here is a daft theory about how you got the joke—and it is probably not quite fair to Jerry Fodor and other language-of-thought fans, but they haven't offered any alternatives: In come some sentences (exactly the sentences written above), through the eyes of those who read the joke and the ears of those who hear it. Their arrival provokes a mechanism that goes hunting for all the relevant sentences—all those on our list—and soon brings them into a common workspace, where a resolution theorem prover takes over, filling in all the gaps by logical inference.

Some such sententialist theory of cognitive processing is the direction in which Fodor has gestured, but nobody has produced a plausible version. No one believes the theory sketch just given, we trust, but even if it and all its near kin (the other sententialist/inference engine theories) are rejected as theories about how you got the joke, our list of propositions is not for that reason otiose or foolish or spurious. It actually does describe cognitive conditions (very abstractly considered) that do have to be met by anyone who gets the joke.

We can easily imagine running the experiments that would prove this. Strike off one belief on that list and see what happens. That is, find some people who lack that belief (but have all the others) and tell them the joke. They will not get it. They *cannot* get it, because each of the beliefs is necessary for comprehension of the story. In other words, we have counterfactual-supporting generalizations of the following form: If you don't believe (have forgotten) that *p*, then you won't get the joke.

Here is an empirical prediction that relies on the scientific probity of talking about this list of beliefs, even though the items on that list do not refer to anything *salient* in the head, but are mere *abstracta:* This joke will soon be extinct, rendered too obsolete to provoke a laugh in a generation or two. Why? Because in this age of wash-and-wear clothing, young people are growing up without ever having seen anybody iron. Some do not know what an iron looks and feels like, and their numbers are growing. For that matter, telephones are changing shape and heft all the time too, so the essential belief in the similarity of shape and heft of telephone receiver and iron is also going to vanish. That belief is not going to be reliably in the belief pool of normal audiences, so they would not get the joke. You would have to explain it to them—"Well, back in the olden days, irons sorta looked and felt like . . . "—and then of course it would no longer be a joke. This example could be multiplied many times over, showing that the power of folk psychology or the intentional stance as a calculus of abstracta is not in the least threatened by the prospect that no version of Fodor's "language of thought" model of belief is sustained by cognitive neuroscience.

Note that introspection leaves open the question of what cognitive mechanisms are involved in getting the joke. Some people may report having had quite detailed imagery while listening to the joke. Others may say they "got it" with only the faintest traces of imagery. In either case, the list of beliefs on which their amusement depended would be surprisingly long. No one has the experience of consciously entertaining such a lengthy set of propositions, even though it must be true that the information expressed by those propositions is somehow in their possession, and has aptly and swiftly enabled them to see the point.

We should not be fooled by the apparent immediacy of the reaction to the joke—or by the absence of any discernible steps between hearing the joke and getting it, or by the impossibility of listing all the propositions required to get the joke—into thinking that those who get the joke must therefore have made a mental jump that did not require any cognitive representation of the intermediate computational steps. Such a viewpoint simply reflects a "fossil trace" of the Cartesian sententialist metaphor, as if propositions must necessarily be processed in their sentential form simply because it is possible and seems natural to us to state them in that form. The neural processes that allow us to get a joke are, by definition, cognitive processes, however nonsentential, emotionally mediated, parallel, and widely distributed in the brain they may be. The problem facing neuroscientists who wish to understand belief is not to discover *how* "belief propositions" are processed by the brain, but rather to discover *if* beliefs are processed in propositional form at all, and if not, to discover what those neural processes which admit of high-level descriptions in terms of propositional processing are actually computing.[3]

We consider it extremely unlikely that beliefs are represented in the brain as sentence-like data structures. Beliefs are ubiquitous guides and influences on our every action and waking reaction, but they are not the familiar items of daily phenomenology. We are unaware of our beliefs, for the most part. We just act on them. Are such "subpropositional" beliefs "memories"? In the fundamental sense defined earlier, yes—for they involve the ability to encode useful information and to decode it in precisely those circumstances where it can be useful. In the stronger, more common sense of the word, however, they are not memories. Any use of the brain's plasticity to store information can be counted as memory in the most basic sense, but we must be careful not to conflate this sense of the word with the everyday concept of memory.

The Intentional Stance

Because we do not want to explain beliefs by first assuming their existence, in the manner of the sententialist theories that assume the existence of proposition-like data structures in the brain, we suggest that the most useful way to think of a belief is as an explanatory tool, rather than as an object in need of explaining. Instead of continuing the attempt to define a belief as an entity that an organism might have or not have (in the concrete, binary-valued way that a library can have a particular book or a poem can have a particular line), a belief must be defined in terms of the circumstance under which a belief could be justifiably *attributed* to that organism. What is meant when it is asserted that an organism has a belief, we propose, is that its behavior can be reliably predicted by ascribing that belief to it—an act of ascription we call taking the intentional stance.

The suggestion is not simply that the adoption of such a definition might be a heuristic for sidestepping the question of what a belief "really" is, but the stronger suggestion that all there is to having a belief that *p* is a system that is efficiently (and, in the strongest cases, most efficiently) predictable under the assumption that it believes that *p*.

This suggestion is intended to carry ontological, rather than simply methodological, weight.

We do not propose to justify this claim in detail here, as much space has already been devoted to that end elsewhere (see Dennett, 1987a, 1988, 1991). Rather, we want to emphasize only that this definition means that we must define a (propositional) belief in the same way that we have defined a memory, as a particular kind of useful knowledge. To say that *x* believes that *p* is to assert that *x*'s behavior (verbal and otherwise) demonstrates a particular kind of regularity; namely, just that kind of regularity which justifies the "projection" (the subjective assumption by the observer) of *x*'s intentionality about *p*. Although there might be many different ways in which (and mechanisms by which) *x* can behave so as to allow for the ascription of intentionality about *p*, we can be sure that the following must be true of all such *x*'s:

1. *x* must "know how" (in a loose sense that includes "be structured so as") to act in order to produce a regularity that allows for the ascription of *p* (and, trivially, must either be able to produce it or be unable to produce it for purely mechanical or external, rather than epistemological, reasons);
2. *x* must be self-motivated (in a loose sense that includes "be structured so as") to produce the regularity described above) (because simply knowing how to state or act in accordance with a belief—or being forced at gunpoint to state or act in accordance with a belief—does not entail believing);
3. the observer (which may be *x* himself) who is making the attribution to *x* of a belief that *p* must be (implicitly or explicitly) satisfied that he is able to discern (independently of whether he actually is able to discern) the regularity that *x*'s behavior demonstrates. In other words, in order to attribute a belief that *p*, an organism must simultaneously attribute to himself (that is, act in a way that seems to him to accord with) the belief "I know what it means to believe that *p*." In order to adopt the intentional stance toward others, one must also adopt it toward oneself.

The third requirement is of particular interest, for a couple of reasons.

The first reason is that the requirement that an organism must be satisfied that it is able to discern the intentionality-defining regularity, is recursively dependent on the very requirements to which it belongs. To be able to discern a regularity in *x*'s behavior is to hold a belief about what it means for *x* to believe that *p*. It is therefore necessary that the attributor fulfill the three conditions outlined above for attributing a belief (in this case, toward himself): he must know how to act consistently with his belief that *x* believes that *p*, must be self-motivated to do so, and must be satisfied that he is able to do so. The recursion is grounded (that is, infinite recursion on the third requirement is avoided) by the fact that the attributor will, by definition, certainly be satisfied that he is able to recognize just those regularities which he himself must be satisfied that he is able to recognize. This is not to say that we must know everything that we believe, which is not necessarily true, but only that, in order to satisfy the above three requirements for believing any proposition *p*, we must also satisfy those requirements for believing another proposition: namely, the proposition that we know what it means to believe that *p*.

The second reason that the third requirement is of particular interest is because it states that the intentional stance can be grounded in an organism's own satisfaction with its ability to detect a relevant regularity, rather than its actual (objectively measurable) ability to detect any such regularity. The existence of an implicit or explicit subjective satisfaction that an organism can detect an appropriate regularity with respect to a proposition *p* (the existence of a belief that it knows what it means to believe that *p*) in the absence of any empirical supporting evidence is not a rare or unlikely event: jealous lovers who unjustly accuse their mates of infidelity, sports fans who treat their home team as if it had a consistent personality across changes in its membership, and people who believe political campaign promises, among many others in our world, all exhibit signs that they have projected belief onto a system that can be demonstrated to be undeserving of the honor. In a less forgiving world than the one we humans have managed to create for ourselves, such errors would be little tolerated. Natural selection would act as the quick and final judge of which systems for recognizing intentional patterns had sufficient practical utility. In our human world, however, it is possible to mix in a great many idiosyncratically grounded intentional projection systems with the few that are required by us all—hence the great variance in the beliefs that human beings are willing to agree they hold.

Note that the claim that beliefs depend on an organism's ability to self-assess its own beliefs should not be interpreted as supporting a radical epistemological relativism, since it is evident that the most important means by which human beings assess their satisfaction with their beliefs is by being satisfied that they hold the beliefs that society taught them how be satisfied that they held.[4] Children raised to believe that *p* are much more likely to grow up believing that *p* than children raised to believe that not-*p,* a fact which has been the source of considerable distress in the course of recorded human history among the more adamant proponents of not-*p*.

Many of the objections raised against the view of belief offered by the intentional stance spring from a desire once identified by Ludwig Wittgenstein as a "need": the desire to maintain a folk psychological category of belief, which views belief as some sort of mysterious essence inherent in the entity to which the belief is attributed. Wittgenstein (1983) wrote: "If you talk about essence—you are merely noting a convention. But here one would like to retort: there is no greater difference than that between a proposition about the depth of the essence and one about—a mere convention. But what if I reply: to the depth that we see in the essence there corresponds the deep need for a convention" (p. 65).

Acceptance of the view of belief offered by the intentional stance means that a wide variety of objectively different phenomena (including those that may demonstrably be modulated by different brain regions within a single species) may justifiably be called belief. The demand for a definition which defines belief in terms that are independent of situational contingencies is ill founded, in the same way that the demand for a detailed but context-free definition of evolutionary fitness is unfounded. In both cases, the only definitions that can be provided must either be so general as to be unhelpful as an explanatory device in any specific situation (for instance, "an organism is fit because it is well suited for living in its environment") or so specific that no wide generalization of it will be possible (we should not expect to learn much about why swallows have survived as a species from a detailed understanding of the eating and mating habits of a species of jellyfish).

We would like to discourage the demand for a context-independent definition of belief, and encourage the idea that the definition of a belief in any particular circumstance is equivalent to the identification of the contingencies which allow that belief to be attributable. (Language, of course, allows us to produce behaviors that make it very easy to attribute beliefs to ourselves and to each other.) In just the same way that the biologist who wishes to discuss "fitness" must identify the relevant constraints in both the species under discussion and the environment in which its members are competing for the chance to reproduce, we suggest that the psychologist who wishes to discuss "belief" with scientific precision must identify both the environment in which the belief under discussion is attributed (that is, the relevant rules of interpretation under which the attribution is made) and the actions of the organism that fall under those rules of interpretation.[5]

It might be further clarifying to compare the current status of beliefs in scientific terminology with the status of genes prior to Crick and Watson's identification of DNA as the vehicle of genetic transmission. At that point in time, genes had already been identified by their functional roles as characterized in the (over-) simplifications of transmission genetics. They were "whatever they are" that could play those relatively well-defined roles. The attempt to fit old-style gene talk into the realities of molecular biology required some large adjustments in the understanding of what a gene might be, to the point that serious controversy exists about whether it is right to talk about genes at all.

Although the situation with beliefs is much less sanguine, the dimly imagined hope that beliefs may turn out to be like what genes were once thought to be—salient structures, identifiable (with minimal procrustean adjustments) as the vehicles of previously well-defined content elements—presupposes what is simply not true: that there is a relatively rigorous, precise, predictive science of beliefs and their functional interactions. This is the myth of "East Pole" Cognitive Science (see Dennett, 1987b, 1998) and the "physical symbol system" hypothesis, which has not yet been borne out at all. We can keep this fond hope in mind without succumbing to the imperative that urges us to consider it an inevitable development.

Conclusion

In terms of Monod's dialectical view of the history of philosophy, we have offered definitions of memory and belief that fall rather closer to the side of the Scruffies than that of the Neats. Although it may be dismaying or frustrating to some, the history of the concepts clearly demonstrates that neater definitions are likely to be philosophically problematical. The complex network of implicit and explicit knowledge that underlies the categories of both "belief" and "memory" rests on the ability of that network to both define and recognize its own coherence. Belief and memory thus fall under the definition of knowledge that the Italian anti-Cartesian philosopher Giambattista Vico (1710/1988) gave nearly three centuries ago: the process of making the objects of the mind (or, as we might say today, the apparent objects of the mind) correspond to each other in shapely proportion.[6] Because organisms may have developed numerous different methods for making the apparent objects of the mind correspond to one another in shapely proportion—and because the methods by which such shapeliness may be detected are informed by many different kinds of learning—we must be wary of any definition that treats memory and belief as if they could be simple primitives existing only inside the brain. They are rather summary descriptions, imposed from the top down, of a set of diverse and complex mechanisms that are all attempting to achieve a single common end: to keep the useful facts of history from going immediately inert.

References

Aristotle. (c. 350 B.C./1941). On memory and reminiscence. In *The basic works of Aristotle* (J. I. Beare, Trans.). New York: Random House.

Bartlett, F. C. (1932). *Remembering.* Cambridge: Cambridge University Press.

Broad, C. D. (1925/1962). *The mind and its place in nature.* London: Routledge and Kegan Paul.

Charniak, E. (1974). Toward a model of children's story comprehension. Unpublished. MIT lab report 266.

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

Dennett, D. (1987b). The logical geography of computational approaches: A view from the East Pole. In M. Brand and M. Harnish (Eds.), *Problems in the representation of knowledge.* Tucson: University of Arizona Press.

Dennett, D. (1988). Précis of *The intentional stance. Behavioral and Brain Sciences, 11,* 495–546.

Dennett, D. (1991). Real patterns. *Journal of Philosophy, 87,* 27–51.

Dennett, D. (1995). Do animals have beliefs? In H. L. Roitblat and J-A Meyer (Eds.), *Comparative approaches to cognitive science.* Cambridge, MA: MIT Press.

Dennett, D. (1996). *Kinds of minds.* New York: Basic Books.

Dennett, D. (1998). *Brainchildren: Essays on designing minds.* Cambridge, MA: MIT Press.

Efran, J., Lukens, M., and Lukens, R. (1990). *Language, structure, and change.* New York: W. W. Norton.

Goldman, A. I. (1986). *Epistemology and cognition.* Cambridge, MA: Harvard University Press.

Holland, R. F. (1954). The empiricist theory of memory. *Mind, 63,* 464–468.

Hume, D. (1739/1896). *Treatise of human nature* (L. A. Selby-Bigge, Ed.). Oxford: Clarendon Press.

James, W. (1890/1950). *Principles of psychology,* Vol. 1. New York: Dover.

Locke, J. (1700/1979). *An essay concerning human understanding,* (P. N. Nidditch, Ed.). Oxford: Clarendon Press.

Malcolm, N. (1963). *Knowledge and certainty: Essays and lectures.* Englewood Cliffs, NJ: Prentice-Hall.

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

Maturana, H. R. (1970). Biology of cognition. In H. R. Maturana and F. J. Varela. (1980). *Autopoesis and cognition: The realization of the living* (pp. 1–58). Dordrecht, The Netherlands: D. Reidel.

Mill, J. (1869). *Analysis of phenomena of the human mind,* Vol. 1. London: Long, Green, Reader, and Dyer.

Monod, J. (1974). *Chance and necessity* London: Fontana.

Plato (347 B.C./1928). *The works of Plato* (I. Edman, Ed.). New York: Tudor.

Plotkin, H. (1994). *Darwin machines and the nature of knowledge.* Cambridge, MA: Harvard University Press.

Reid, T. (1815/1969). *Essays on the intellectual powers of man.* Cambridge, MA: MIT Press.

Russell, B. (1921). *The analysis of mind.* London: Allen and Unwin.

Ryle, G. (1949). *The concept of mind.* London: Penguin Books.

Smith, B. C. (1996). *On the origin of objects.* Cambridge, MA: MIT Press.

Vico, G. (1710/1988). *On the most ancient wisdom of the Italians unearthed from the origins of the Latin language* (L. M. Palmer, Trans.). Ithaca: Cornell University Press.

Wiener, N. (1948/1991). *Cybernetics.* Cambridge, MA: MIT Press.

Wittgenstein, L. (1958). *Philosophical investigations.* Oxford: Blackwell.

Wittgenstein, L. (1983). *Remarks on the foundations of mathematics.* Cambridge, MA: MIT Press.

Zechmeister, E. B., and Nyberg, S. E. (1982). *Human memory: An introduction to research and theory.* Monterey, CA: Brooks/Cole.

Notes

1. Notwithstanding his common association with the representative theorists, Locke's view of memory was actually more subtle than this metaphor suggests. He goes on to note that "our Ideas being nothing, but actual Perceptions in the Mind, which cease to be anything, when there is no perception of them, this laying up of our Ideas in the Repository of Memory, signifies no more but this, that the Mind has a Power, in many cases, to revive Perceptions, which it has once had . . . And in this Sense it is, that our Ideas are said to be in our Memories, when indeed, they are actually no where, but only there is an ability in the Mind, when it will, to revive them again" (p. 150). This passage suggests that Locke did not think of the memory simply as a static storehouse of images, but rather as a dynamic ability to evoke or reconstruct an image, a metaphor more closely in keeping with modern scientific views of memory, which is close to the view of memory we champion here.
2. This example is drawn from Dennett, 1987, pp. 76–77.
3. David Marr (1982) made a relevant distinction between what he (rather misleadingly) called the computational level and a lower level he referred to as the algorithmic level. We might prefer to call his computational level the functional level, concerned as it is with what function, in an abstract and idealized sense, the system under consideration is intended to compute: that is, concerned as it is with competence rather than performance. The lower algorithmic level is concerned with the abstract details of how that function is actually implemented by the system. (A third, lowest level—the hardware level—is concerned with the concrete engineering details of the implementation of that particular algorithm.) Marr's point was that mistakes at the highest (computational) level made it difficult or im-

possible to adequately address questions at the lower levels, because errors at the computational level made it difficult or impossible to distinguish between informative fact and artifact at the lower levels. Looking for propositions in the brain might be very much like looking for perfect earth-centered circles in the orbits of the heavenly bodies. The high-level requirements of the theory mislead one into blurring a distinction we decidedly do not wish to blur: the one between noise (theoretically neutral facts relating to performance) and data (theoretically important facts relating to competence).

4. As Efran, Lukens, and Lukens (1990, p. 111) put it in discussing the closely related phenomena of meaning: "Although we understand (and agree with) the notion that meanings are social constructions, that does not imply that they can or should be modified at will. Participating in a culture is a commitment to abide by established language conventions. Capricious renaming of actions for short-term gain may entail unexpected, hidden, long-term risks. If you cheapen the verbal coin of the realm now, it is hard to escape the inflationary effects later."
5. In actual practice, of course, the rules of interpretation that seem to be commonly accepted in the company in which the attribution of beliefs takes place, can often be simply assumed by the attributor to be accepted by the intended audience of his intentional attribution. If they are not, he is communicating with the wrong audience.
6. This view is reflected in the etymology of "belief." According to the Oxford English Dictionary, the word derives from a degraded form of the original Teutonic "galaubian," which means "to hold estimable or pleasing; to be satisfied with," intensified by the addition of the prefix "be." Thus, etymologically, a belief is something with which one is thoroughly satisfied or much pleased.

Memory, Brain, and Belief

Edited by
Daniel L. Schacter
Elaine Scarry

Harvard University Press
Cambridge, Massachusetts
London, England / 2000

Copyright © 2000 by the President and Fellows of Harvard College
All rights reserved

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Memory, brain, and belief / edited by Daniel L. Schacter and Elaine Scarry.
p. cm.
Includes bibliographical references and index.
ISBN 0-674-00061-7 (alk. paper)
1. Memory Congresses. 2. Belief and doubt Congresses.
3. Cognitive neuroscience Congresses. I. Schacter, Daniel L.
[DNLM: 1. Memory Disorders—physiopathology Congresses. 2. Brain
Congresses. 3. Delusions Congresses. 4. Knowledge Congresses.
5. Memory—physiology Congresses. 6. Self Concept Congresses. WM
173.7 M53253 2000]
QP406.M44 2000
612.8′2—dc21
DNLM/DLC
for Library of Congress 99-40552