DANIEL C. DENNETT

HOW TO STUDY HUMAN CONSCIOUSNESS EMPIRICALLY OR NOTHING COMES TO MIND

Consciousness is making a comeback in psychology, but there is still residual skepticism, anxiety, and confusion about how to approach this perilous phenomenon scientifically. There are those, after all, such as Thomas Nagel, who have argued that approaching consciousness must inevitably involve leaving science behind – the objective world of science recedes as we close in on the subjective world of consciousness, on *what it's like to be* a conscious being. There may in the end be something to this suspicion, but I will try to show how little it could come to. It is perfectly possible to study consciousness, carefully conceived, empirically.

I will support this claim by describing a method, and giving its rationale. Versions of the method are familiar in experimental psychology, though it is never, I think, practiced with attention to quite the set of principles and constraints I will describe. The method has close kin in the history of philosophy and psychology. There is little that is new in it; it looks a bit like Wundtian introspectionism, a bit like Husserlian – or better, Schutzian – phenomenology, and even like Quine's imagined exercises in "radical translation".

A hallmark of the method is its cageyness, its metaphysical minimalism; it begins by cautiously saying nothing at all about what consciousness might be, or even where it might be found. We all know, of course, one place consciousness might be found: in *us*, the more or less normal, intelligent adult human beings. And moreover we all are quite sure that the reason – or at least a major reason – why we are so sure *we* are conscious is because we have told each other, in one way or another. That is, we all think that one can learn *something* about the nature of what others are conscious of – what it is like to be them – from listening to what they tell us. For just these reasons, then, I will restrict the domain of the method to just such

Synthese **53** (1982) 159–180. 0039-7857/82/0532-0159 $02.20

apparently normal human beings, and more particularly to what they say, but the method itself requires no assumptions – for the moment at least – about whether any or all such organisms really are conscious, and if so, of what. We start with a class of subjects picked out of the crowd by those rough and ready pretheoretical assumptions but decline to advance any premature theses about this class. For all we know, some of the members of this class may be masqueraders – zombies without a glimmer of consciousness in them.

It is often the case in experimental psychology that groups of such subjects are placed, one at a time usually, in rather constrained experimental situations where they are asked to perform various intellectual tasks, solve problems and puzzles, make judgments, and so forth. One feature of such experimental situations is that they are – or can easily be made – objectively describable in many of their relevant details. For instance, we can adopt a resolutely third-personal perspective, viewing the subjects *from the outside only*, as it were. We can restrict ourselves to such data-gathering as tape-recording, videotaping, the timing of button pushing, measurement of brain waves, galvanic skin response and so forth. If we are scrupulous about these matters, then whatever it is that we are studying, and however well or ill we are studying it we will be studying it empirically.

Let us focus first on the record of noises – vocal noises mainly – made by subjects (and experimenters) during some experiment or group of experiments. Now one *could* try to be a physicist or physiologist and devise a purely physical or physiological theory or model of the subjects that would explain and predict these noises, but if one fact is now abundantly clear from the last century of such endeavors, it won't work. It might work in principle – in principle one might also predict exactly where and when lightning will strike – but in fact there isn't a prayer. But happily there are other levels of analysis at which we might construct good, predictive theories. For instance, if we can turn our recording of the noise stream into a transcript, a text composed of words of the subject's language, there are all sorts of well-attested regularities, dependencies, and redundancies to exploit. This first step yields a radical reconception of the data, an abstraction from acoustic and physical properties to strings of words (though still adorned with precise temporal features, and perhaps other clearly physical features as well). The physics and physiology of speech production and perception are still poorly understood in spite of vigorous and brilliant research, so disappointingly little can be said at this point about the regular relations that must exist between the occurrence of words (or phonemes) and physical features of the acoustic signal, but there is nevertheless sufficient interpersonal agreement in the transcription process to ensure that a transcript is still a reliably objective refinement of the data. After all, any two typists with good hearing and a good vocabulary will independently produce transcripts that agree on all but a tiny percentage of the words. Many inaudible or garbled words can be extrapolated by appeal to obvious assumptions of grammar or the subject's manifest intentions, and such imponderable matters of judgment or interpretation that remain can be isolated, marked, and if need be discarded as data.

So far, I take it the method described is cut-and-dried and uncontroversial. It yields what I will call a *text* – a purified text, literary scholars might call it – and it is not quite *given*, since the process of transcription is also, as just noted, a process of (obvious) interpretation, which depends on assumptions about the language being spoken and on some of the subject's intentions. We can be systematically cautious about the latter assumptions, restricting ourselves to what we might call assumptions of grammatical and lexical intentions. Thus we would boldly purify "from reft to light" to "from left to right", and insert a definite article at the obvious place in "... as soon as the car turns corner, ..." but we would leave "I seem to hear an orange blur" in unadulterated inscrutability, pending our next step.

Moving beyond the text is a step in which the caution begins to pay big dividends. On the one hand, we must now venture into the controversial and treacherous territory of *hermeneutics*, the interpretation of the text. On the other hand, we can really be quite confident that the text we find produced by our subjects consists of speech acts; not mere pronunciations or recitations but assertions, answers to questions, comments, self-corrections, requests for clarification. We are generally prepared to assume, that is, that the text is a product of a process that has an *intentional interpretation*: it consists of things the subjects *wanted to say*, of *propositions* they meant to *assert*, for various *reasons*. In fact of course, some reliance on such assumptions was playing a background role in our previous step, purifying the text. (Why would anyone *want to say* "from reft to light"?) The reliance on an intentional interpretation of the subjects is in any event an ineliminable part of such experiments, both in the interpretation of the data, and in the prior process of experimental design. Whatever dangers we run by adopting the intentional stance, they are the price we must pay for gaining access to a host of highly reliable truisms we wish to exploit in the design of experiments. For instance, there are many reasons for wanting to say things, and it will be convenient to exclude some of these by experimental design. Sometimes people want to say things not because they believe them but because they think their audience wants to hear them. We take the obvious steps to diminish the likelihood that this desire is present or effective: we tell our subjects that what we want to hear is *whatever they believe*, and we take care not to let them know what it is we hope they believe. We do what we can, then, to put them in a situation where, given the *desires* we have inculcated in them, they will have no better option than to try to say what they in fact *believe.*

Another application of the intentional stance towards our subjects is required if we are to avail ourselves of such useful event-types as button-pushing. Typically, pushing a button is a way of performing some conventionally fixed speech act, such as *asserting that* the two seen figures appear superimposed to me right *now*, or answering that *yes*, my hurried, snap judgment (since you have told me that speed is of the essence) is that the word that I have just heard was on the list I heard a little while ago. For many experimental purposes, then, we will want to unpack the meaning of these button pushings and incorporate them as elements of the text. Which speech act a particular button pushing can be taken to execute depends critically on the intentional interpretation of the interactions between subject and experimenter that prepared the subject for the experiment.[1] Steps can be taken, and are routinely taken, to remove sources of ambiguity and uncertainty in the experimental situation, so that *one* intentional interpretation of the text (button-pushings and the like included) is overwhelmingly dictated – has no plausible rivals.

The upshot of this rehearsal of the obvious is that we can and routinely do prepare subjects of whom it can be assumed with confidence

(1) that their noises (and other conventional acts) are interpretable as text

(2) that their text is interpretable (defeasibly, and to a first approximation, as we shall see) as a sincere and reliable (error-corrected) account of their current *beliefs* or *opinions*.[2]

Obvious though these results may seem at first glance, do they not already contradict my earlier claim that the method would not make any assumptions about the consciousness of such subjects? One might hold them to do that, on the grounds that only a conscious being can be held to have beliefs, and that performing speech acts is something only a conscious creature can do, but equally one *might* refrain from assuming there was any such necessary relation between consciousness and intentional interpretation, even between consciousness and the correct interpretation of some behavior as the performance of speech acts. That is what I choose to do. Since the prospect of zombies behaviorally indistinguishable from conscious beings is often advanced as a skeptical challenge to functionalistic theories of the mind such as my own (and such as the theories that typically are seen to undergird such fields as cognitive psychology), I shall adopt the tactic of conceding the possibility, and then point out that nothing in the method as so far described will uncover such a zombie amidst the conscious subjects, and hence zombiehood presents no obstacles to the process of text purification and interpretation so far described. This observation thus challenges the presumption of a necessary connection between intentional interpretation and consciousness; *if* there is such a necessary relationship, then, since some of our subjects might, for all we know so far, be zombies, we had better not speak of assumptions about the subjects' *beliefs* and *desires* and *speech acts*, but only of their apparent beliefs, desires and speech acts. From now on, then, when I speak of beliefs, desires, speech acts, and other intentional states and actions, understand my terms to be surrounded with cautious scare-quotes.

Happily for us, there is an analogy at hand to help us remember what we are and are not assuming at this point in our investigation. Consider the particular sub-branch of hermeneutics that deals with the interpretation of fiction. Some texts, such as novels and short stories, are known – or assumed in any event – to be fictions, but this does not stand in the way of their interpretation, but rather frees the hermeneuticist from certain distractions (about such difficult topics as truth, reference, and sincerity). This permits the flat dismissal of certain sorts of questions, and at least a crisp isolation and postponement of other difficult questions. I want now to consider some fairly familiar facts about the projects and principles of this area of literary criticism. It may even be a case of the Humanities lending a Helping Hand to Science.

Consider the semantics of fiction.[3] A novel tells a story, but not a true story, except *per accidens*. In spite of our knowledge or assumption that the story told is not true, we can, and do, speak of what is *true in the story*. "We can truly say that Sherlock Holmes lived in Baker Street and that he liked to show off his mental powers. We cannot truly say that he was a devoted family man, or that he worked in close cooperation with the police" (Lewis, p. 37). What is true in the story is much, much more than what is explicitly asserted in the text, of course. It is true that there are no jet planes in Holmes' London (though this is not asserted explicitly or even logically implied in the text) but also true that there are piano tuners (though – as I recall – none are mentioned, or, again, logically implied). In addition to what is true and false in the story, there is a large indeterminate area: while it is true that Holmes and Watson took the 11:10 from Waterloo Station to Aldershot one summer's day, it is neither true nor false that that day was a Wednesday ('The Crooked Man'). Lewis (and others) have a panoply of clever technical proposals to make about how to handle the formal semantics of fiction without embarrassment, but these proposals will not concern me directly, for while I am fascinated by them, I am not worried by the problems they are designed to solve. That is, perhaps some people are deeply perplexed about the metaphysical status of fictional people and objects, but not I. In my cheerful optimism I don't suppose there is any deep philosophical problem about the way we should respond, ontologically, to the results of fiction; fiction is *fiction*; there *is no* Sherlock Holmes. There are delicious philosophical problems about how to say (strictly) all the things we unperplexedly want to say when we talk about fiction, but as I say, these will not be my concern.

What I want to draw attention to is the fact that the interpretation of fiction, the fleshing out of the story, the exploration, if you will, of "the world of Sherlock Holmes", is undeniably do-able, with certain uncontroversial results. First, the exercise is not pointless or idle; one can learn a great deal *indirectly* about a novel, about its text, about its point, about the author, about the real world, by learning about *the world portrayed* by the novel. Second, if we are cautious about identifying and excluding what are clearly judgments of taste or preference (e.g. "Watson is a boring prig") we can amass a great volume of unchallengeably objective fact about the world portrayed. All interpreters agree that Holmes was smarter than Watson; in crashing obviousness lies objectivity.

Third – and this fact is a great relief to students – knowledge of the world portrayed by a novel can be independent of knowledge of the text of the novel. I could probably write a passing term paper on *Madame Bovary*, but I've never read the novel – even in English translation. I've seen the BBC television series, so I *know the story*, I know what happens in that world. Facts about the world of a fiction are (of course) purely *semantic level* facts about that fiction; they are independent of the syntactical facts about the text (if the fiction *is* a text). We can compare the film *West Side Story* with *Romeo and Juliet*; by describing similarities and difference in what happens in those worlds, we see similarities in the works of art that are not describable in the terms appropriate to the syntactical or textual (let alone physical) description of the concrete instantiations of the fictions. One can describe *what* is represent*ed* in *Madame Bovary* independently of *how* the represent*ing* is accomplished.[4] (Typically, of course, one doesn't try for this separation, and mixes commentary on the world portrayed with commentary on the author's means of accomplishing the portrayal, but the separation is possible.) One can even imagine knowing enough about a world portrayed to be able to identify the author of a fiction, in entire ignorance of the text or anything purporting to be a faithful translation. Learning indirectly what happens in a fiction we might be prepared to insist: only Wodehouse could have invented that preposterous misadventure. We think we can identify sorts of events and circumstances (and not merely sorts of *descriptions* of events and circumstances) as Kafkaesque, and we are prepared to declare characters to be pure Shakespeare. Many of these plausible convictions are no doubt mistaken, but not all of them. I mention them just to illustrate how much one might be able to glean just from *what is represented*, in spite of having scant knowledge of *how the representing* is accomplished.

Now let us apply the analogy to the problem facing the experimenter who wants to interpret the texts produced by his subjects. Let us consider the advantages of adopting the tactic of interpreting these texts *as fictions* of a sort, not as literature of course, but as generators of a theorist's fiction. Just as the literary critic or hermeneuticist of fiction lets the text *constitute* a (fictional) world, a world determined by fiat by the text, exhaustively extrapolated as far as extrapolation will go and indeterminate beyond, so our experimenter, whom I shall now call the heterophenomenologist, lets the subject's text constitute what I shall call the subject's heterophenomenological world. We can thus postpone the knotty problems about what the relation might be between that world with its fictional denizens, and the real world. Note of course that the literary trick that inspires us here need not be restricted to works *intended* as fiction by their authors; we can describe a certain biographer's Queen Victoria, or the world of Henry Kissinger, with blithe disregard of the author's presumed intentions to be telling the truth, and referring noncoincidentally to real people, living or dead.

One obvious advantage – but not by any means the main one – of playing the heterophenomenological game is that one can thereby remain officially neutral about the zombie problem. Consider a novel written in the autobiographical mode, the tale of a fictional narrator. Just as the literary interpreter may describe for us the character of the fictional narrator of such a novel (the fictional "referent" of all the "I's" and "me's" in the text – Holden Caulfield, for instance, or Ishmael) without addressing the question of whether or not this narrator is *really the author*, partly the author, or based on the author or any other real person, so the heterophenomenologist can compose detailed descriptions of the *heterophenomenological subject*, the logically constituted referent (or fictional "referent") of all the "I's" and "me's" in the text produced by some (apparently) human subject, while postponing the question of whether there really is any ego or subject or soul *in there*, as it were, for those pronouns to refer to. The particular *point of view* of the subject is an objectively extrapolatable abstraction about which much can be said independently of whether or not that point of view is, as we might say, *inhabited*. If our subject is a zombie, the point of view we define and describe by the heterophenomenological method is vacant, not occupied – there's nobody home, you might say – but curiously enough this dire prospect (well, isn't it a dire prospect?) has no apparent bearing on the utility for theory of the construct. It will be perspicuous, no doubt, to describe much of what goes on in and with the experimental subject (the *thing* we are investigating) from the subject's point of view (from the heterophenomenological subject's point of view), whether or not there really is – as there appears to be – something it is like to be the subject in question.

I am pleased to discover this tactical ploy made available by the analogy between heterophenomenology and the exegesis of fiction, and I shall make more of it later, but I view it as a by-product of the theoretically more useful suggestions to be found in the analogy. Once we have settled upon whatever portion of a subject's heterophenomenological world interests us, we must face the tremendously difficult question of what the relation is *or could be* between the events, objects, denizens of that world, and the events, processes, objects, and states *inside the subject's body* (in the brain, we generally suppose). Here we find guidance once again in the world of literary theory, but this time in a particularly embattled part of that world. There are beleaguered factions in the literary world who want to ask, and answer, questions such as the following. Upon what real person in the author's acquaintance is this character modeled? Is this character really the author's mother in disguise? What real events in the author's childhood have been transmogrified in this fictional episode? What awful memory drove the author to create this episode? What is the author *really* trying to say, or do? Then one can enter the fray on whether *asking the author* would be of any use in settling such questions – supposing one ought to ask him or her in the first place. There are those who insist that the author's opinions, candidly expressed or not, on such matters are well-nigh useless, and those who find that dogma well-nigh incredible. There are those who think such issues as the psychoanalysis of the author, and even the unadorned biography of the author, are utterly irrelevant to the only questions that are appropriately posed about the object of Art. These controversies, so superficially portrayed by me here, nevertheless provide insights into the novel problems of interpretation facing the experimental student of consciousness.

Whether or not such questions of biographical source are irrelevant to Art, they surely compose a perfectly possible inquiry that often yields results. Maybe one *shouldn't* read novels as thinly veiled autobiographies of their authors, but one often *can*, thereby discovering a great deal about the author's life – more, often, than the author ever dreamt he was revealing. Sometimes, in fact, it can be very plausibly argued that an author, wanting (perhaps unconsciously) to tell the world about some remarkable event or person in his life, has been forced (perhaps unwittingly) to express himself *allegorically*, in effect. The only expressive resouces available – for whatever reason – to this author did not permit a direct, factual, unmetaphorical, recounting of the events he wished to recount; the fiction he composed is the compromise or net effect; it may be *drastically reinterpreted* (if necessary, over the author's anguished protests) to reveal a true tale, about real people and real events. Since, one may sometimes argue, it is surely no coincidence that such and such a fictional character has these traits, we may reinterpret the text that portrays this character in such a way that its terms can then be seen to refer – in genuine, non-fictional reference – to the traits, and actions of a real person. Portraying (fictional) Molly as a slut may quite properly be seen as slandering real Polly, for all that talk about Molly[5] is *really* about Polly. The author's protestations to the contrary may convince us, rightly or wrongly, that the slander is not, in any event, a conscious or deliberate slander, but we have been persuaded at last by Freud and others that authors, like the rest of us, are often quite in the dark about the deeper wellsprings of their (artistic) intentions. They often don't know why they want to say what they want to say. If there can be unconscious slander, there must be unwitting reference to go along with it.

When we turn back, once again, to the interpretation of the texts of subjects in experiments, we know (or assume) they are doing the best they can to tell the truth, but might it not be that when it comes to fixing the *real reference* (if any) of the terms subjects use when portraying their heterophenomenological worlds for us, we might on occasion be justified in similarly usurping the authority of the author and reinterpreting texts to yield (sometimes true, sometimes reliable) accounts of "deep" or in any event inaccessible events and processes occurring within the heads of the subjects? For instance, suppose our subject describes a manipulation of mental images he has just performed. That is to say, the subject yields a text, which, by scrupulous heterophenomenological extrapolation, we get to yield a portrayal of a heterophenomenological world in which various (fictional) things called mental images cavort in various ways. We theorists on the outside can *talk about these images*, give them names, describe their careers, just as readily as the critic can recount the adventures of Alyosha and Ivan.[6] Having obtained a clear view of these heterophenomenological objects, we then ask: but what – if anything – is our subject *really* talking about? Is there something really going on in him which behaves in such ways as to render it no coincidence that these are the ways he describes his mental images? Can we reinterpret his text to refer to those internal events? Perhaps the subject's limited expressive resources are such that this is the only way he can put the matter – even to himself.[7]

This suggestion might benefit from a concrete, if partly science-fictional,[8] example of what I mean. A few years ago, there was a robot at SRI in Palo Alto named Shakey.[9] Shakey was a sort of box on wheels with a television eye and instead of carrying his brain around with him, he was linked to it (a large stationary computer) by radio – a dubious arrangement in my experience. Shakey lived indoors in a few rooms in which the only other objects were a few boxes, pyramids, and wedges. One could communicate with Shakey via a computer terminal, in a severely restricted vocabulary of semi-English. "PUSH THE BOX OFF THE PLATFORM" would send Shakey out, finding the box, locating a ramp so he could push the ramp into position so he could roll up the ramp onto the platform, and push the box off. Now how did Shakey do this? How, in particular, did Shakey distinguish boxes from pyramids with the aid of his television eye?

The answer, in outline, was readily apparent to observers, who could watch the process happen on a screen. The original, grainy, television image of, say, a box would appear on the screen; the image would then be purified and rectified and sharpened in various ways, and then, marvelously, the boundaries of the box would be outlined in white – and the entire image turned into a line drawing. Then Shakey would analyze the line drawing; each vertex was identifiable as either an *L* or a *T* or an *X* or an arrow or a *Y*. If a *Y* vertex was discovered, the object had to be a box, not a pyramid; from no vantage point would a pyramid project a *Y* vertex. Shakey had a "line semantics" program for wielding such general rules to determine the category of the object whose image was on the screen. Watching the screen, observers might be expected to suffer a sudden dizziness when it eventually occurred to them that there was something strange going on: they were watching a process of image transformation on a screen, but *Shakey wasn't looking at the screen.* Moreover, Shakey wasn't looking at any other screen on which the same images were being transformed. There were no other screens in the hardware, and for that matter the screen they were watching could be turned off or unplugged without detriment to Shakey's processes of perceptual analysis! Was this screen some kind of fraud? *For whom* was the screen? Only for the observers. What then was its point, and what relation did the events they saw on the screen bear to the events going on inside Shakey?

The screen was for the observers, but the *idea* of the screen was also for the designers of Shakey. How could you get a computer to take a television camera input and yield box-identification as output? The signal coming from the television camera is a stream of zeroes and ones, each representing a cell of light or dark on the "retina" of the camera. Suppose, to oversimplify, the retina was a grid of 10,000 cells: 100 by 100. Then a sequence of 10,000 zeroes and ones would encode a single frame, a timed sampling of the light falling on the retina. Now computers are designed to be fed streams of zeroes and ones, but what could a computer possibly *do* to such a stream to "figure out" that there was a box in front of the camera? Here is where the idea of the screen is valuable. Suppose we spread the stream out on another grid – our screen, in fact – in a hundred rows of a hundred, reading from left to right just like sentences in a book (and unlike commercial TV which does a zigzag scan). Notice that we can now specify, purely in terms of erasing and printing ones and zeroes in the stream, operations that would "purify" the image. We can heighten contrast, removing the salt from the pepper and the pepper from the salt, with operations such as these:

(a) Erase any 1 with 7 or 8 adjacent 0s and print 0; erase any 0 with 7 or 8 adjacent 1s and print 1

(Two-dimensional adjacency can be defined in terms of the position in the long sequence: the digits adjacent, in this sense, to the digit in position 374 are the digits in positions 273, 274 (just "above" 374) 275, 373 (just to the "left" of 374) 375, 473, 474, 475.)

Then we can locate, and outline in white, the vertical light-dark boundaries by scanning for sequences of 1s followed by sequences of

0s, followed precisely 100 positions later by another sequence of 1s followed by a sequence of 0s, and so forth. Once such a boundary has been located, it can be turned into a crisp vertical white line by judicious erasing and printing so that something like 001100 occurs exactly every hundred positions in the sequence. Slopes and horizontal boundaries can be similarly found and outlined. That gives us the line drawing, and now we need to define templating operations that will similarly locate and distinguish the different varieties of vertex. A harder bit of mathematical design, but you can see roughly how to proceed.

This final process yields a "judgment" about the geometrical category of the object being scanned. That is how Shakey tells the boxes from the pyramids. Shakey's brain is a computer, and more precisely a von Neumann machine, an elaboration of Turing's classical idea of a Turing machine. It is a *linear*, sequential computer, doing one thing at a time. In this instance, we can quite directly envisage its operations in classical Turing machine terms: the printing and erasing of 0s and 1s in an input sequence, but however the hardware is configured, the *spatial* properties of its operations are irrelevant or nearly so. (I suspect some *recherché* topological properties might inevitably be shared by the working hardware and the changing shapes on the screen.) Even though there is no actual image locatable in 3-space in the hardware, the operations in the hardware are strictly homomorphic to the events we observed on the screen – and those were genuinely imagistic: of a particular size, color, location and orientation. So in *one* strict, literal sense, Shakey does *not* detect boxes by a process of image transformation; the last real image in the process is the one that is focussed on the receptive field of the camera. In another strict but metaphorical sense, Shakey *does* detect boxes by a process of image transformation. That this second sense is in some regards metaphorical can be brought out by noting the variety of questions one would expect to be answerable about images that are not answerable about the "images" processed by Shakey's hardware. They have no color, no size, no orientation, no clear location.

Suppose now we indulge in a little science fiction, and cross Shakey with another famous character in artificial intelligence circles: Terry Winograd's SHRDLU,[10] who manipulates (imaginary) blocks and then answers questions about what it is doing and why. An interchange with our new version of Shakey, redesigned to include a more sophisticated repertoire of verbal actions, might go like this:

Why did you move the ramp?
SO I COULD ROLL UP ON THE PLATFORM,
And why did you want to do that?
TO PUSH THE BOX OFF.
And why did you want to do that?
BECAUSE YOU TOLD ME TO.

But suppose we ask Shakey:

How do you tell the boxes from the pyramids?

What should we design Shakey to want to say in reply? Here are three possibilities:

(1) I scan each 10,000-digit-long sequence of 0s and 1s from my camera, looking for certain patterns of sequences, such as . . . blahblahblah (a *very* long answer, if we let Shakey go into the details).

(2) I find the light-dark boundaries and draw white lines around them in my mind's eye; then I look at the vertices; if I find a *Y*-vertex, for instance, I know I have a box.

(3) I don't know; some things just look boxy or cubical to me; it's a sort of intuition, or gestalt, or boxy raw feel or something. It just comes to me.

Which is the right sort of thing for Shakey to say? I suggest that all three answers get at a version of the truth, and which Shakey would say depends on what *access* we design Shakey's expressive capacity to have to his perceptual processes. Perhaps there will be good reasons (of engineering, say) to deny deep (detailed, time-consuming) access to the intermediate perceptual analysis processes – the processes that ultimately govern Shakey's speech act productions.[11] But whatever communicative and self-descriptive capacities we endow Shakey with, there will be a limit to the depth and detail of his expressible "knowledge" of what is going on in him, what he is doing. If the best answer he can give is a type (3) answer, then he is in the same position with regard to the question of how he tells pyramids from boxes that *we* are in when asked how we tell the word "sun" from the word "shun"; we don't know how we do it; one sounds like "sun" and the other sounds like "shun" – that's the best we can do. If Shakey can respond as in (2), there will still be other questions he cannot answer, such as "And how do you draw white lines on your mental images?" Suppose we design Shakey to (want to) give (2)-type answers to our questions about his perceptual processes. Shakey says he processes images. Unbeknownst to him, we unplug and throw away the screen. Are we then entitled to tell him that we know better? He isn't really processing images, though he thinks he is? If he were a realistic simulation of a person he might well retort that we were in no position to tell *him* what was going on in his own mind! *He* knew what he was doing! If he were more sophisticated, he might grant that what he was doing might only be allegorically describable as image processing – though he felt overwhelmingly inclined so to describe what was happening. Of course if we are diabolical, we can rig Shakey to have entirely spurious ways of talking about what he is doing – to want to say things about what is going on in him that have no truth-preserving interpretation at all; Shakey is just confabulating.

And that, finally, is the major reason for going to the roundabout trouble of treating heterophenomenology as analogous to the interpretation of *fiction*. It is beginning to emerge, from a wide variety of experiments, that people are often just wrong about what they are doing and how they are doing it.[12] It is not that they *lie* in the experimental situation, but that they confabulate; they make up likely sounding tales without realizing they are doing it; they fill in the gaps, guess, speculate, mistake theorizing for observing. They are, then, unwitting creators of fiction, but of course to say they are unwitting is to grant that what they say is, or can be, an account of *exactly how it seems to them*. They tell *what it is like* to them to solve the problem, make the decision, and since they are sincere (apparently) we are prepared to grant that that is – must be – what it is like to them, but then what it is like to them turns out to be a poor guide to what is really going on in them.

In a recent exchange in *Psychological Review* this issue is presented as a disagreement over the reliability of verbal reports as data for cognitive psychological theory. In 'Telling More than We Know: Verbal Reports on Mental Processes', (*Psychological Review*, 1977), Nisbett and Wilson summarize a host of experiments and studies that apparently unmask human subjects as inveterate confabulators about their thinking. The shocking, almost paradoxical conclusion seems to be that we don't even know our own thinking! We have, in Gunderson's nice phrase, *underprivileged access* to our own mentation.[13] In a recent rebutting paper, 'Verbal Reports as Data', Anders Ericsson and

Herbert Simon argue that although there are indeed many circumstances under which subjects' verbal accounts of their own information processing are highly dubious and unreliable, there is a model available (which is consistent, by the way, with the less detailed model I sketched in 'Towards a Cognitive Theory of Consciousness')[14] that permits the theorist to distinguish between the circumstances in which verbal reports will be reliable and the circumstances in which they won't. While there is much in these papers that repays a philosopher's attention, and deserves philosophical comment, I will restrict myself here to one point. Both Nisbett and Wilson and Ericsson and Simon are shy about talking about consciousness. As is typical in cognitive psychology, it is almost as if there were a tacit agreement that human subjects were in fact talking zombies whose talk was being treated as symptoms of internal processing to be assessed for reliability.

Since those internal processes are what the cognitive psychologists are interested in studying, this makes perfect sense.[15] If we ask, "But what about consciousness?" they can candidly reply without embarrassment that they are ignoring consciousness, but they needn't say that. The heterophenomenological method permits them instead to say: we construct portions of the subject's heterophenomenological world, and then our question is: when and why do the things that happen in that world tell us the truth about the things that happen in the subjects' brains? The heterophenomological world we construct from the subject's verbal reports is an objective, *outsider's* view of that subject's consciousness (if he or she *is* conscious, of course!). It is guaranteed to be accurate because we can put it to the subject for corroboration; we can close the loop and permit the subject to revise, adjust, disavow, confirm, embellish, edit the text, producing new chapters *ad lib*, until the heterophenomenological world portrayed asymptotes in convergence with the subject's autophenomenological world (if there is one). Of course if our subject is a zombie, then this feedback loop just leads to what would better be described as either stabilization, or endless elaboration, of the merely heterophenomenological world of that zombie.

It is time to take stock of this examination of heterophenomenology before turning to autophenomenology and its particular mysteries. First, I must rush to issue a *caveat* about my account of Shakey. I do *not* mean to suggest that Shakey is a realistic model of human perception, belief or (with his SHRDLU attachments) speech production. The system as described is only the crudest of sketches of the sorts of complex relationships that would have to exist in any good psychological model of these capacities. And in particular, I don't want to be understood as suggesting that the way in which Shakey's actual computing processes can be viewed *metaphorically* as image-processing is just like the way the brain's actual processes might be metaphorically described as image-processing. Brains may be computers, and hence in a mathematically powerful but mechanistically superficial sense, equivalent to Turing machines, but they surely do not have the machine architecture of a von Neumann machine. The example of Shakey was merely meant to illustrate in *one* way a larger possibility opened up for theorists to compose theories that are imagistic at one level of description, but not *all the way down.*

At the outset I spoke of the metaphysical minimalism of the heterophenomenological method. This is what I meant: the heterophenomenologist describes a world, the subject's heterophenomenological world, in which there are various *objects*, in which *things happen*, but if we ask "What *are* these objects, and what are they *made of*"? the answer is "Nothing"! What is Mr. Pickwick made of? Nothing. Mr. Pickwick is a fictional object, and so are the objects described, named, mentioned by the heterophenomenologist. The heterophenomenologist takes himself – at the outset – to be speaking about nothing, but as we know from the example of literary interpretation, this can be an activity that is neither unprincipled nor pointless. Sometimes one can express useful and illuminating facts by speaking about things that are fictional.

There is another way in which heterophenomenology is metaphysically minimal, or better: scientifically minimal. While it purports to be a way of characterizing the relationship between language and consciousness – that particular sort of consciousness that our pretheoretical intuitions and traditions suppose to be intimately connected with the capacity for language – it does this while being almost entirely non-committal about the actual nature, structure, and real properties of whatever-it-is we take ourselves to be talking about when we tell others how it is with us, what it is like to be us. That is, while this view treats of a phenomenon that is dependent upon the text-producing capacity of some organisms, it does not presuppose that this real phenomenon is somehow itself *linguistic*, made out of words or sentences in the head, for instance. The public text produced by subjects, the text recorded and transcribed, is, of course, made out of words, but the heterophenomenologist is neutral with regard to the relation between this public text and the private (just in the sense of *internal*) represent*ing* which, we might say, it partially co-represents. *Partial co-representation* is the relation that an English translation bears to the original French text of *Madame Bovary*, but it is also the relationship that either of these texts bears to the film or videotape of *Madame Bovary*. When two represent*ings* represent the same (fictional) world, they are co-representations. Probably no two different representings – especially in different media – can portray exactly the same world; hence I speak of partial co-representation.

Now some think that when we speak, our words of natural language are a sort of translation of sentences in our language of thought, but another possibility is that the relation of those public words to our private thoughts is rather more like the relation between those recent *novelizations* and the original films they are parasitic upon. No doubt there are other, better possibilities. John Maynard Keynes was once asked whether he thought in words or pictures. His reply, which the heterophenomenologist applauds, was "I think in thoughts".[16] Finding out what *they* might be is the *next* task, which the heterophenomenologist can attack from a starting point of studied neutrality. So far, his characterization of what happens in (heterophenomenological) consciousness is purely at the semantic level; it is an account of what is represent*ed* and tells us nothing *yet* about the structure or substance of the represent*ing*. This is not to say that the heterophenomenologist must remain forever neutral on this score; the hope, in fact, is that if one can just get a clear, detailed and well-confirmed description of what is represented, this will force constraints on hypotheses about how the representing must be done. I think Roger Shepard's brilliant experiments on mental imagery are best viewed in this light.[17] In showing how surprisingly rich and *imagistic* the heterophenomenology of some subjects – good imagers – is, he drives up the requirements on the representing machinery. Modest hypotheses about that machinery that looked plausible before his experimental explorations of the heterophenomenology of good imagers are now seen to be inadequate.

A strikingly counterintuitive way of characterizing this side of heterophenomenology's minimalism is to note that it is a "black box" psychology *par excellence* – like behaviorism! It does *not* hypothesize inner mechanisms at all, but achieves its organizational and predictive power by an *indirect* characterization of input-output relations – relations between all the publicly accessible variables we record and manipulate in the experimental situation. It does not, however, pretend to be the whole story – as some earlier black box theories unwisely purported to be – but rather a valuable prolegomenon to the whole story, a data-organizing phase of the whole scientific enterprise.

Now, finally, what of autophenomenology? Does the heterophenomenological enterprise, whatever its utility to science, simply leave the *real* problems of consciousness untouched? John Searle, in rebutting my commentary on his attack on "strong" artificial intelligence,[18] explicitly warns you not to let me hoodwink you with this *hetero* approach. "Remember", he admonishes, "in these discussions, always insist on the first person point of view. The first step in the operationalist sleight of hand occurs when we try to figure out how we would *know* what it would be like for others" (p. 451). I guess you should have walked out at the beginning of my talk, if Searle is right, or shouted me down, but it is too late for that. Now have I tricked you? Why would I want to do a thing like that?

Let us see if heterophenomenology is unfair to autophenomenology. First of all, as we have already noted, when you are put in the heterophenomenologist's clutches, *you get the last word.* You get to edit, revise and disavow *ad lib*, and what you insist upon is granted constitutive authority to determine what happens in your heterophenomenological world. You're the novelist, and what you say goes. What more could you want? Surely you know your own mind? That is to say, there is a lot of your mind that you *don't* know, as we are now discovering, but that is – by definition really – the *un*conscious part. The part you do know is the part you can tell us about. So tell us. We'll trust you. We *won't* trust you to tell us the truth about the processes occurring in you, and if you think you are authoritative about them, you should think again, for no one has that sort of God-like infallibility to report on what is actually happening. But we do trust you to tell us just how it *seems* to you; we *constitute* you as an authority on that.[19] Of course if there are any cleverly designed zombies around, who aren't *really* conscious, but whose unconscious text-producing capacities are so marvelously sophisticated as to create the illusion of a conscious subject, we will be taken in, and concede consciousness, with all its rights and privileges, to some undeserving nonentities.

But now, finally, aren't we all a little old to be believing in zombies? What is it that these supposed zombies lack that we lucky ones enjoy? A soul, perhaps; an ego, a self. A something-it-is-like-something-to-be. A locus of meaning, understanding, and value. Taking the last point first, why wouldn't a zombie be a locus of value? Why wouldn't a zombie be a member in good standing of the class of things with *interests*, with desires to be satisfied, projects to complete, harms to be protected against? Even the lowly lobster, however zombie-like we may suppose him to be, is cunningly organized to take self-regarding steps to prolong its own existence. When distributing good and ill, then, may not the utilitarian (for instance) count the lobster as a suitable, if modest, receptacle for some portion? Why not the zombie, then? Let us not be racists or speciesists. Some of your best friends may be zombies.

I have been playing along with this zombie idea – for tactical reasons that should now be obvious – but in fact I think (in case you have not already guessed) that the concept is just incoherent. The idea of a being that could pass all the heterophenomenological tests but still be a merely unconscious automaton strikes me as simply bizarre. I don't know how to argue against it, however, beyond presenting the case I have just presented for heterophenomenology. This leaves a symmetrical standoff, however, for those who think the secret light of consciousness is untouched by my reflections are equally unforthcoming in support of their creed.

What little more I can offer might best be considered to be sympathetic – if unasked for – therapy. Part of your problem, you who remain unpersuaded, is this: when I announce that the objects of heterophenomenology are mere theorist's fictions, you are tempted to pounce on this and say, "That's *just* what distinguishes the objects of autophenomenology from the objects of heterophenomenology. *My* autophenomenological objects are perfectly *real* – though I haven't a clue what to say they are made of. When I tell you, sincerely, that I am rotating a mental image, or imagining a purple cow, I am not just unconsciously producing a word-string to that effect, cunningly contrived to coincide with some faintly analogous physical happening in my brain; I am consciously and deliberately reporting the existence of something that is *really there*! It is no mere theorist's fiction *to me*! I see it with my own eyes ... well, no, I see it in my *mind's* eye".

But reflect more cautiously on this speech. You are not just unconsciously producing a word-string, you say. Well, you *are* unconsciously producing a word-string; you haven't a clue to how you do that, or to what goes into its production. But, you insist, you know *why* you're doing it; you *understand* the word-string, and *mean* it. I agree, but merely point out that understanding and believing a sentence heard (or heard in one's mind's ear) is *not* a matter of using the sentence as a sort of mental movie-projector which, when understanding is achieved, *produces* a mental object, or *displays* a mental scene. That is a tempting but hopelessly wrong idea. Once one banishes it, the apparently striking difference between the objects of heterophenomenology and the objects of autophenomenology fades. Raskolnikov's dark brown hair, like the purple flank of the cow you imagine, does not exist. Consciousness is not a *process* that *makes things*; it is a state of being informed – or misinformed – about what is actually happening.

Tufts University

NOTES

[1] Not all button-pushing consists in speech acts, of course. Some may be make-believe shooting, or make-believe steering-rocket firing, for instance.

[2] In 'How to Change Your Mind' in *Brainstorms: Philosophical Essays on Mind and Psychology* (Montgomery, Ut. Bradford Books, 1978) I adopt a conventional use of "opinion" that permits me to draw a sharp distinction between beliefs proper and other more language-infected states (which I call opinions). While I shall not presuppose familiarity with, or acceptance of, that distinction here, I do mean what I say here to be about both beliefs and what I call opinions in *Brainstorms*.

[3] There have been several very interesting articles on this topic recently. I am drawing particularly on David Lewis's, 'Truth in Fiction' *American Philosophical Quarterly* **15** (1978): 37–46. See also Kendall Walton, 'Pictures and Make Believe', *Philosophical Review* **82** (1973): 283–319; 'Fearing Fiction', *Journal of Philosophy* **75** (1978): 5–27; and Robert Howell, 'Fictional Objects: How They Are and How they Aren't,' in D. F. Gustafson and B. L. Tapscott, eds., *Body, Mind and Method* (Dordrecht, Holland: D. Reidel Publishing Co., 1979), pp. 241–94.

[4] I discuss this in more detail in 'Beyond Belief' in Andrew Woodfield, ed., *Thought and Object*, (Oxford University Press, 1981).

[5] "Molly-about", in Nelson Goodman's sense of Pickwick-about. See his 'About', *Mind* **71** (1961): 1–24.

[6] See my 'Two Approaches to Mental Images', in *Brainstorms* for an earlier version of these claims.

[7] On the reasons why an intelligent creature's expressive or representational powers must be limited, see Douglas Hofstadter, *Gödel, Escher, Bach, an Eternal Golden Braid,* (New York: Basic Books, 1979).

[8] Because I shall both oversimplify and embellish the description in the interests of clarity and vividness.

[9] See, e.g., Bertram Raphael's account of Shakey (of which he was one of the creators) in *The Thinking Computer: Mind Inside Matter,* Freeman, 1976. Today Shakey, minus his computer brain, sits, like Jeremy Bentham, in Nils Nilssen's office at SRI, where I paid him a sentimental visit in 1980.

[10] Terry Winograd, *Understanding Natural Language* (New York: Academic Press, 1972). SHRDLU is discussed in Ch. 7 of *Brainstorms,* and in 'Beyond Belief' (*loc. cit.*).

[11] See K. Anders Ericsson and Herbert Simon, 'Verbal Reports as Data', *Psych. Review,* **87** (1980): 215–50.

[12] See, e.g., R. Nisbett and T. DeC. Wilson, 'Telling More than We Know: Verbal Reports on Mental Processes', *Psych. Review,* **84** (1977): 231–59, and M. Gazzaniga and J. Ledoux, *The Integrated Mind* (New York: Plenum, 1978).

[13] Keith Gunderson, 'Asymmetries and Mind-Body Perplexities', in David M. Rosenthal, ed., *Materialism and the Mind-Body Problem* (Englewood Cliffs, N.J.: Prentice-Hall Inc., 1971) p. 117. Gunderson seems to understand something slightly different by his use of the phrase.

[14] In *Brainstorms,* Ch. 9, reprinted from C. W. Savage, ed., *Perception and Cognition: Issues in the Foundations of Psychology* (Minneapolis: University of Minnesota Press, 1978).

[15] I argue that the proper domain of cognitive psychology is such internal processes and *not* the beliefs and desires of folk psychology in 'Three Kinds of Intentional Psychology', in R. Healey, ed., *Reduction, Time and Reality,* (Cambridge University Press, 1981).

[16] Reported to me by Isaiah Berlin, in conversation.

[17] See Roger N. Shepard and Lynn A. Cooper, *Mental Images and their Transformations* (Bradford Books, forthcoming) for an overview and discussion of these results.

[18] See his 'Author's Reply' to the critics of his 'Minds, Brains, and Programs', in *Behavioral and Brain Sciences* (September, 1980): 417–58.

[19] Not quite an *incorrigible* authority. While your word is and must be in general the best source of information on how it seems to you, it is possible for you to be wrong even about this. See Raymond Smullyan, 'An Epistemological Nightmare', and the Reflections following it, in Douglas R. Hofstadter and Daniel C. Dennett, eds., *The Mind's I* (New York: Basic Books, 1981).