SYNTHESE LIBRARY

STUDIES IN EPISTEMOLOGY,
LOGIC, METHODOLOGY, AND PHILOSOPHY OF SCIENCE

Managing Editor:

JAAKKO HINTIKKA, *Florida State University, Tallahassee*

Editors:

DONALD DAVIDSON, *University of California, Berkeley*
GABRIËL NUCHELMANS, *University of Leyden*
WESLEY C. SALMON, *University of Pittsburgh*

VOLUME 194

PERSPECTIVES ON MIND

Edited by

HERBERT R. OTTO
Department of Philosophy, Plymouth State College (USNH)

and

JAMES A. TUEDIO
Department of Philosophy, California State University, Stanislaus

D. REIDEL PUBLISHING COMPANY

A MEMBER OF THE KLUWER ACADEMIC PUBLISHERS GROUP

DORDRECHT / BOSTON / LANCASTER / TOKYO

bility of the thesis that consciousness plays a crucial, global role in mental processing? Supporters of McIntyre's phenomenological view would reply that Rey's strategy overlooks the very feature of mental processing that makes qualia and meaning possible. Are these counterarguments convincing? Suppose a robot were capable of human-like responses in problematic situations. Would it make any functional difference to the robot whether it was capable of entertaining qualitative experiences? Would it make any defensible difference to *us*? How could we even determine whether it had such a capacity? What test might we employ to settle the issue?

Rey proposed one kind of test--computational programs that function in accordance with "rational regularities" mimicking human mental operations. Were they to react to problem-solving situations in ways indistinguishable from our own, would it matter whether the robot was experiencing qualia? Rey avoids this question, for he apparently assumes that the concept of "qualia" is every bit as suspect as the concept of "consciousness" when it comes to developing an ontology of mental processing. But since there is serious doubt whether Rey's program would in fact capture the essential features of mental processing, we must not shy from the issue: to what extent *would* such a mechanism exhibit a capacity to entertain qualitative experience? If not at all, would we be justified in concluding that the robot lacked a crucial structure of mental processing? Or would it be more appropriate to conclude that the qualia issue is *irrelevant* to questions concerning the nature and structure of such processing?

In the next paper, James H. Moor proposes and analyzes strategies for testing computational mechanisms for evidence of qualitative experience. Professor Moor argues that such tests can never be decisive. It is impossible, he concludes, to justify the claim that a robot could entertain qualitative experiences "functionally analogous" to those we experience. But it would be a mistake, he adds, to make a great deal of the issue when assessing the design of robots. For if a robot exhibits functionally analogous behavior, then Moor sees nothing to be gained from "testing" whether or not the mental processes of the robot manifest a level of subjective awareness. Moor maintains that research should focus instead on determining which mental operations are associated with our behaviors *when we are talking of* experiencing qualia. It will suffice to design machines capable of simulating these operations. If it turns out that robots behave as we do, we will find it impossible to prove--although difficult not to believe--that such mechanisms are entertaining subjective experiences like our own. From this perspective, success or failure of the AI/cognitive science enterprise turns, not on successful production of qualitative experience in robots, but on the degree to which computational mechanisms exhibit problem-solving behavior analogous to our own in those situations where, from our point of view, experience of quality seems so important.

JAMES H. MOOR

Testing Robots for Qualia

1. The Meat/Metal Distinction

A computer recently electrocuted its owner just after he had purchased a another computer. Did the computer kill from jealousy? Was it seeking revenge? Such explanations are fun to give, but I assume that nobody takes them seriously. Among other things, the behavior of today's computers is not sophisticated enough to even begin to convince us that they actually have qualitative experiences such as emotions and feelings.

Moreover, any attempt to give emotions and feelings to a computer by adding some affective behavior seems superficial. Imagine a good chess-playing computer enhanced to display emotion. The superior chess-playing computer might emit a synthetic chortle during a game when a human opponent made a particularly stupid move. Such a computer might gloat after winning a game by saying something like "nice try for a human." If it lost, the computer might have a temper tantrum. But these particular enhancements make the computer more obnoxious than feeling. "User unfriendly" computers are no more emotional than "user friendly" ones. Such behavior may arouse our feelings; but it is not really an expression of the computer's feelings. Common sense tells us that behind the facade of behavior there is still emptiness. A computer is emotionally hollow--void of feeling.

Perhaps a more promising approach to constructing a computer with qualitative experiences is to base its design on the internal workings of a human being. For the sake of argument, let us suppose a researcher conducts an extensive study of the human brain and related chemistry, and he becomes thoroughly knowledgeable about how the brain functions. Certain complex systems of the brain are understood to be responsible for certain feelings and emotions and for producing particular behavior patterns. Suppose the functionality of these systems is meticulously duplicated in computer hardware. Wetware is converted to dryware. The functionality of the systems, the relationship of inputs and outputs, is maintained although the makeup of the systems is changed. It will be useful to connect the inner systems to outer motor and sensory systems which are also essentially computer circuitry. The result of this endeavor is a robot that has an electronic brain which is functionally analogous to a human brain and has peripheral devices which are analogous to human motor and sensory systems. Now if the researcher has done the job properly, the robot should act in the world much the way a human being does. The robot should see with artificial eyes and grasp with artificial hands. From time to time the robot should show feelings. If its hand is squeezed too hard, it should

H. R. Otto and J. A. Tuedio (eds.), Perspectives on Mind, 107–118.

react accordingly.

But, does such a robot really have qualitative experiences? Does it really have sensations, feelings, and emotions? According to a functionalist view of a mind, the answer is "yes". Functionalism is not so much a single theory as a constellation of theories which share a central notion that a mind can be understood as a complex functional system. (Putnam, 1960; Fodor, 1968/1981). On the standard functionalist interpretation the components of a functional system can be realized in many ways both biologically and nonbiologically. On this view, humans are computers that happen to be made out of meat. Of course, it is also possible on this analysis of mind for a computer made out of electronic components to have a full mental life.

I think many people remain skeptical about the potential inner life of a robot, because even if a robot behaves in sophisticated ways, it seems to be made out of the wrong stuff to have qualitative experiences. Paul Ziff, who denies that robots can have feelings, puts the point in the following way:

> When clothed and masked they may be virtually indistinguishable from men in practically all respects: in appearance, in movement, in the utterances they utter, and so forth. Thus except for the masks any ordinary man would take them to be ordinary men. Not suspecting they were robots nothing about them would make him suspect.
>
> But unmasked the robots are to be seen in all their metallic lustre. (1964, p. 99)

How important is the meat/metal distinction with regard to having feelings and emotions? Is biology crucial for a mental life? It seems possible, if not likely, that a system which is functionally equivalent to a human being but made out of nonbiological parts may behave as if it had a mind, but in fact have no subjective experiences at all. This, I take it, is the point of the standard "absent qualia" objection to functionalism. If a functional theory doesn't capture qualia, i.e., our qualitative experiences, then it is an inadequate theory of mind. (Block, 1980a,b)

In this paper I want to examine some tests and arguments which are designed to resolve the issue of whether an electronic robot is made out of the wrong stuff to have qualitative experiences such as sensations, feelings and emotions. I will assume the robot under discussion behaves in a manner closely approximating human behavior and that it has an internal organization which is functionally equivalent to relevant biological systems in a human being.

2. *The Transmission Test*

One approach to gathering nonbehavioral evidence for robotic experience is to tap into the inner processes of a robot's brain and to transmit the results. The transmission can be either indirect or direct. With indirect transmission the robot's inner processes are connected via a transmission link to a display board which can be examined by our sensory systems. The display board contains output devices which reveal what the robot senses. A television screen shows us what the robot sees, a speaker let's us hear what the robot hears, and so on. Various dials indicate the levels of emotional states.

If such a test were actually run, I think we would be skeptical about the results. Suppose there is an area on the display board which allows us to feel what the robot feels with its artificial hand. The robot touches a hot piece of metal with its hand, and we in turn touch the appropriate place on the display board. The board feels warm to us. Does the robot really feel the warmth? Or does the display board get warm merely as the result of a straightforward causal chain which is triggered by a hot piece of metal contacting the robot's hand?

Even if the information on the display board were sophisticated, as it would be with a television picture, I don't believe we would regard it as a reliable indicator of what the robot actually experiences. Some years ago there was a robot built at Stanford called Shakey. Shakey rolled about several rooms plotting pathways in order to travel from one place to another. Shakey had a television camera which allowed it to compute its position relative to other objects. Researchers could watch a television set to see what Shakey saw. But did Shakey see anything? Shakey used some of the information from the television camera input, but why should we believe Shakey actually experienced anything? Television cameras are transmitters of information, not experiencers. Thus, evidence about inner experiences gathered from a television display is not convincing. The television camera in Shakey could have been mounted on a cardboard box and still have transmitted the same robust pictures.

There are two clear shortcomings of this indirect transmission test. First, the evidence gathered by us is limited to information on the display board. The information on the display board may be so abstracted from the nature of actual experiences that it will not be persuasive. For example, evidence for emotional states in the form of dial readings is less convincing than ordinary emotional behavior itself. Second, the information picked up by our sense organs may tell us nothing more than the state of the display board. What we want to know about are the *actual experiences* of the robot. The display board output is determined by causal processes, but this causal chain may not reflect the robot's experiences. A videotape machine provides a nice display on a television but the

videotape machine itself presumably has no qualia.

The transmission test can be improved, however, by eliminating the display board and transmitting the information in the robot's brain directly to a human brain. James Culbertson has proposed an experiment along this line. In his words, "The way to show that the machine is sentient, i.e., experiencing sensations, percepts, and/or mental images, is to connect it to the nervous system of a human observer." (1982, p. 6)

Suppose we set up a direct transmission test in which the analogous portions of the robot's brain are connected with a transmission link directly to a human brain. Now we can imagine that when the robot touches a piece of hot metal, the human in the test experiences what he would experience if he had touched a piece of hot metal. The experiences passed to the human monitor in the direct transmission test are not limited to sensory experiences. Presumably, emotional information can be passed on as well. If the robot feels angry, then the human monitor will feel anger. Hence, in the direct transmission test the subjective experiences of the robot can be experienced directly by a human being.

But is this direct transmission test really a good test? Perhaps it is an improvement over the indirect version, but the fundmental difficulty which lurks behind the indirect version lurks behind the direct version as well. Is it the case that the human monitor experiences what the robot experiences? Or, is it the case that the robotic apparatus simply generates experiences in a human? The human monitor has experiences initiated by the transmission link connected to the robot, but a human subject would also have experiences if connected to any machine generating similar signals.

The situation is not unlike actual results of brain probes on human subjects. Electrodes are used sometimes to stimulate various regions of a patient's brain and the patient reports having various kinds of experiences. In this situation, nobody maintains that the electrodes along with the associated electronic devices are actually having the experiences which are then transmitted to a human. Rather, the explanation is that the electrodes, when properly used, activate neural mechanisms which generate experiences in a human. Perhaps, this is all that is happening in the direct transmission test.

The problem with the transmission test is similar to a problem which confronted John Locke. Locke had to explain which of our ideas really represent external reality and which are largely a product of our mind when influenced by external reality. The Lockean problem *vis-à-vis* the transmission test is to distinguish information which represents an internal reality of a robot from information which is largely a product of our own mind when influenced by signals from the transmission link.

The issue comes down to this. If we are already convinced that entities made out of electronic components can have qualia, then the transmission test seems well-grounded. We are actually tapping into a robot's experiences. But, if we are not convinced that a robot can have qualia, then the transmission test has little force. The robotic apparatus is viewed as a device which generates experiences in us but not one which has experiences of its own. Moreover, a negative result in the transmission test is not conclusive either. A functionalist, for example, can argue that a negative result shows only the inadequacy of the transmission link.

The key to our Lockean predicament is to get rid of the transmission link altogether. We must devise an experiment that does not involve transmission so that we can determine even more directly whether or not a device made out of electronic components can have qualia. Let's consider a thought experiment which allows us immediate access.

3. *The Replacement Test*

Suppose that our robot's brain isn't modelled on just any human brain but on Sally's brain. After the electronic brain is constructed, Sally suffers some brain damage. Suppose further that the damaged portions are critical to her pain system. Sally now feels no pain. Because pain provides important warning of injury, Sally would like to have the damage repaired. Biological repair is not possible but an electronic solution is. Scientists decide to remove the analogous portions of the robot's electronic brain and install them in Sally's brain in order to restore her pain system to normal functioning.

Installation of electronic devices in humans is not farfetched. Electronic devices are now implanted in the nervous system to block unwanted pain signals. Pacemakers are electronic implants that regulate heart function. Other implants in humans regulate and release chemicals. In our experiment a portion of the electronic brain is implanted along with an interface mechanism which permits the normal biological activity to interact with the electronic mechanism.

After replacement of the damaged portion of Sally's brain with the electronic analog, we are in a position to test the mental result of the replacement. We touch Sally's foot with a pin and we ask Sally, "Do you feel pain?" Let's suppose Sally replies, "No, I feel nothing." We move the pin to other locations and insert it somewhat deeper. Each time Sally denies feeling pain. Would such a result show that Sally, equipped with the computer implant, did not feel pain? It is unclear. The assumption of the replacement experiment is that the computer component is a functional equivalent of the portion of the brain being replaced. But, the present evidence would suggest that this assumption was not satisfied. That is, behavioral evidence that Sally is not feeling pain is equally evidence that the computer implant is not functioning properly. If the computer implant

were truly functionally equivalent, then the output from it to the speech center would be such that Sally would say that she *did* feel pain. In other words, if Sally had an intact brain which functioned properly, then the information going into the pain center indicating pain in her foot would be followed by information leaving the pain center enroute to the speech center, and finally resulting in Sally saying "ouch" or at least being put into a state such that when asked about pain, Sally acknowledges pain. An *adequate* computer replacement must do the same thing.

So, let us suppose the computer implant in Sally's head is adjusted or replaced with another computer component so that the correct functionality *is* achieved. Once this is done, Sally responds normally when pricked with a pin. She says, "Ouch!" and readily acknowledges pain when asked.

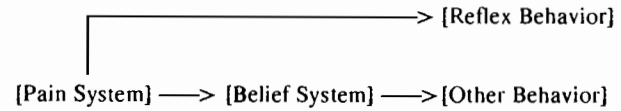
But, suppose we are not convinced by Sally's report of her pain. We ask Sally, "Does the pain feel the same as the pain you felt when your brain was functioning normally and you received such a pin prick on the foot?" Sally might say that it does, but suppose she says, "No." Sally claims that she feels something, but it is quite different from the way it used to feel when she was pricked on the foot. What does such a result show? The evidence indicates that Sally feels something but it isn't quite the normal feeling of pain. The evidence for abnormality of the feeling can, thus, be taken as evidence for readjusting the functionality of the computer implant so that the report of the pain experience is a report of normal pain feeling. In other words, if the replacement were really a functional equivalent of the original brain pain center, then the information sent to the memory areas in the brain should duplicate the information that would have been sent by a normal functioning brain pain center. Because this is not the case in light of Sally's assertion that her current feeling is much different from her old feeling of pain, some functional adjustment is again needed. After the adjustment is made, Sally readily tells us that not only does she feel pain, but the pain is just the same as the pain she used to feel with her brain intact.

As we can see from the foregoing, a difficulty with the replacement test is that the evidence gathered for or against computer feelings is still essentially indirect. Behavioral evidence indicating a lack of feeling is equally evidence for the improper functioning of the computer implant. In principle, the behavioral evidence can always be manipulated by adjusting the functionality of the computer replacement component. This makes the test inconclusive. Appropriate behavior may be the result of massaging the evidence, and thus not indicative of inner feelings at all. Perhaps the test can be made more direct. Rather than a third person report, what is needed is a *first person* report.

So in an attempt at a direct version of the replacement test, a gallant researcher decides to have the electronic components implanted in himself. Now he will know immediately whether or not he feels pain and

whether the pain is just like the pain he used to feel. Of course, he already knows on the basis of the indirect replacement test that his outward behavior will indicate pain, hence others will think he is in pain whether he is actually in pain or not. Indeed, there will be no way for him to signal to others about the nature of his inner experiences. It will do little good to prearrange an eye signal, for instance, where three quick blinks of the right eye means "Trust my behavior; I'm really in pain" and three quick blinks of the left eye means "Ignore my behavior; I feel no pain." Eye blinks are behavior, and if the electronic components are functioning properly then the scientist should give the trustworthy signal. If he didn't give the right signal, then his fellow scientists would know the implant was not functioning properly and would make the appropriate adjustments so that he did give the right signal in the future.

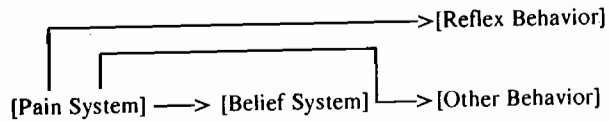
Is this direct version of the replacement test decisive at least for one person? I think not. It seems obvious that our behavior is highly dependent on our beliefs. Thus, it may not be possible for our guinea pig scientist to believe he is not in pain and behave completely as if he is. In other words, for the scientist to act consistently as if he is in pain it will be necessary to implant a functional unit that gives him the belief he is in pain. A highly schematic functional configuration looks like this:



Hence, even if the computer implant does not generate the feeling of pain, the scientist himself will delusionally believe that it does. He will be sincere, though mistaken, in his reports that he feels pain which is similar to the pain he use to feel when his brain was exclusively biological.

My skepticism about the direct replacement test is based on a hypothesis about the way our brain works, viz., that our beliefs are instantiated some way in our brains and as such they play a critical causal role in determining our behavior. Of course, an actor can produce pain behavior without believing he is in pain. But an actor can break character; he can choose not to display the pain behavior. This is not the case in the direct replacement test. Our assumption here is that the behavior of the guinea pig scientist is the same as the behavior of someone who really is in pain. It is hard to imagine such consistent pain behavior occurring without the appropriate belief in pain. I am not suggesting, as I think Shoemaker (1975) does, that an adequate pain belief requires the feeling of pain. My hypothesis is the empirical claim that for a complete

and convincing repertoire of pain behavior the agent must believe he is in pain. I think my empirical hypothesis is reasonable, but it may be wrong--and even if it is right, there may be the possibility that the belief system can be bypassed with regard to the behavioral output, yet receive information directly from the pain system. The functional arrangement would be this:



Can the direct replacement test be relied on in this set up? This arrangement does look more favorable, for now it is not required that the belief system contain the belief that the scientist is in pain in order to causally generate the appropriate pain behavior. Of course, the scientist with the implant will still say he is in pain and act as if he is in pain in the appropriate circumstances, but there is now the possibility that he will believe he is not in pain. It now seems that the scientist can know from a first person point of view whether or not the computer replacement really gives him the subjective feeling of pain. But, because his behavioral repertoire is disconnected from his belief system, he will not be able to relay information to his fellow scientists about the results of the test. The situation may seem strange indeed if he believes that he is not in pain but finds his body acting as if he is in pain. For example, he may observe his own body uncontrollably writing articles about the success of the implant and how it really generates the experience of pain while knowing all along that the implant is a fraud and doesn't do anything except generate the outward appearances of pain.

Even within this functional configuration there may be difficulties. Some outputs of the pain system will become inputs to the belief system. The designers of the computer replacement for the pain system can determine what the guinea pig scientist will believe in the direct replacement test by their choice of outputs of the computer implant. Thus, the designers of the computer replacement can guarantee what the guinea pig scientist will believe about his subjective experiences whether he has them or not. What this suggests is that even this special version of the direct replacement test is not decisive. The guinea pig scientist's belief about his qualia will be both uncertain and ineffable.

In summary, the transmission test and the replacement test are inconclusive. If these tests are viewed through a functionalist framework, then either the right results can be guaranteed or the wrong results can be explained away. Whether robots really have qualia is a contingent matter which cannot be rigorously empirically tested. I don't believe this

defeats functionalism, but it does suggest that a significant part of the argument for functionalism must be non-evidential.

4. *The Argument for Qualia*

What defense does a functionalist have against the charge that it is possible that robots which instantiate functional systems like the ones humans instantiate lack qualia? This problem is the robot corollary to the problem of other minds. Part of the defense is to address the problem of other minds. After all, it is possible that some humans who instantiate biological systems like the ones we instantiate lack qualia. There are many ways of partitioning the human population, granting qualia to some and not to others. Perhaps all humans have qualia except those of the opposite sex, or those who are born in a country other than one's native land, or those who lived during the 19th-century, or those who are not identical with me. These hypotheses are bizarre but not inconsistent. Why do we reject them? A traditional answer is that these other humans are similar to us (me!). By analogy we (I) grant them qualia. But I think this traditional answer is inadequate. A skeptic has only to argue that though these other humans are similar, they are not similar enough.

A better answer is that the attribution of qualia gives us essential explanatory power. Imagine what it would be like if one seriously denied that some humans had qualia. The most ordinary behavior of members of this group would become virtually incomprehensible. How would we understand the actions and words of an absent-qualia human who on a winter's day came inside shivering and complained at length about the cold? There is an extension, but no inflation, of ontology in granting others qualia similar to our own. And, there is an enormous gain in explanatory power in granting others qualia similar to our own. Good explanations are what determine ontology in this case.

Good explanations determine ontology for robots too. If, as I have been assuming, some robots act in ways which closely approximate human behavior, and their brains are functionally equivalent to human brains, then attribution of qualia to robots will be necessary for a reasonable understanding of their actions. I am not denying that there will be lower level explanations of robot behavior in terms of circuitry any more than I am denying that there will be lower level explanations of human behavior in terms of neurology. (Moor, 1978a) I believe these lower level explanations will be compatible with, but in general not as perspicuous as, higher level explanations of behavior in terms of qualia. Of course, it will always be logically possible that attributing qualia to robots is a mistake. But this is a general feature of induction and not a special problem about qualia. (Moor, 1976)

What becomes of the common sense objection that electronic robots are

made out of the wrong stuff to have qualitative experiences? I think the objection does have some force, but it is important to realize that this force rests on an ambiguity in the objection. On one reading, the objection is about the empirical possibility of constructing robots--on this reading, the claim is that electronic components can never be assembled to produce robots that have the appropriate affective behavior or cannot be organized into a functional system that duplicates the functionality of the human brain. In the abstract, I don't find this version of the objection plausible. Assuming human behavior is directed by the human brain and the brain operates through the firing of neurons which can be described by a computable function, a computer could be designed to instantiate the relevant portions of the biological system. In other words, at the abstract level functionalism seems invincible. But, functionalism ultimately must become a scientific theory which specifies concretely what the relevant functionality of the brain is and how electronic components can perform it. Thus, it is possible that the wrong stuff objection will turn out to be correct for some straightforward technical or empirical reason the import of which would be that the relevant sophisticated functionality simply cannot be created in nonbiological material.

The other interpretation of the wrong stuff objection is conceptual. Its claim is that, even if a robot could be constructed such that it exhibited the appropriate behavior and had the appropriate internal functionality, the robot would not have qualia because it would be made of the wrong material. On this level, I think the objection is no longer a claim that can be decided by gathering further empirical evidence. Tests like the transmission test and the replacement test may be helpful if we already attribute qualia to robots, but they will never be decisive against the absent-qualia objection. What is needed to answer this objection is a conceptual argument about explanatory power. If robots which had the appropriate behavior and functionality were constructed, then the increase in explanatory power would eliminate the meat/metal distinction with regard to qualia. In such a situation, there is no wrong stuff; people and robots are both made of chemicals, and as it turns out, chemicals are the right stuff.

END OF MOOR'S ESSAY

—— ∇ ——

COMMENTARY BY THE EDITORS

A key dimension of Moor's argument is directed against the objection that functionalism fails as a theory of mind because it cannot show how to replicate in robots such qualitative experiences as sensations, feelings, and emotions. But since it seems unlikely that we could ever design *tests* for determining the "qualitative character," if any, inherent in a robot's

functional processes, the objection turns out to be ineffective against a functionalist strategy. The real challenge for functionalism, Moor argues, is to develop a scientific theory capable of specifying in concrete terms the relevant functionality of the brain and how it can be performed by electronic components.

Suppose functionalism were eventually to meet this challenge. According to Moor, it could then appropriate the explanatory "leverage" that comes from attributing qualia to computational mechanisms which have been designed in accordance with its theory of mind. After all, the mechanisms would behave pretty much the way we behave. So, since we gain explanatory leverage with respect to human behavior by invoking references to qualia, why wouldn't we gain--and, indeed, find it useful to do so--the same sort of leverage with respect to robots? Thus, whether or not Moor's robot *actually* entertains qualitative experiences, as long as it acts as though it did, we would be justified in ascribing to it the presence of qualia. And, happily, our ontological commitment would go no further than this, for as Moor observes, "good explanations are what determine ontology." (Moor: this volume, p. 115) Doesn't this imply that the "absent qualia" issue is irrelevant to puzzles concerning the nature and function of mental processing?

On the surface, this appears to be a paradoxical position. For how can the concept of "qualia" provide us with explanatory leverage if we have no interest in the actual relation that might hold between qualia and behavior? The following commentaries suggest that Moor's conclusion may be vulnerable to two lines of criticism. Robert Van Gulick's commentary challenges Moor's conception of the constraints within which functionalism must operate as a science of the mind. He maintains that Moor has not really demonstrated that a functionalist account of the mind's "internal organization" is incapable of explaining the relation between qualia and behavior, and that such an account would therefore be better served by just ignoring the issue.

From a second angle, Henry Johnstone questions whether Moor has analyzed the *only* strategies worth considering when it comes to testing robots for qualia. Do the "transmission" and "replacement" tests really exhaust the list? Johnstone suggests the possibility of a "communication" test, a test which he feels might elicit solid evidence regarding the presence or absence of qualia in computational mechanisms.

Van Gulick's commentary, which comes first, begins with a general criticism of Moor's view. He argues that Moor's primary assumption is incompatible with his subsequent rejection of the metaphysical aspect of the qualia issue. Moor has postulated the existence of a "functional equivalence" between his robot's behavior and our own without having a definitive conception of what such an equivalence would require. How, then, can he use this assumption as the basis for conclusions about the metaphysical side of the qualia issue? "Unless we can say what counts as

playing the same functional role as a qualitative state," Van Gulick writes, "we cannot hope to determine whether non-qualia states could play such roles." (Van Gulick: this volume, p. 120) His analysis of the concept of "functional equivalence" hinges on the distinction between *"what some item does* and *how it does it*," and this leads him to stress the theme of *"psychological equivalence."* (p. 121) For Van Gulick, the key lies in determining "how qualitatively differentiated representations function and how such functions might be realized by underlying causal mechanisms." (p. 122) Given such a theory, he argues, the functionalist might well be in a position not only to address the qualia issue, but also to actually make effective use of Moor's "transmission" and "replacement" tests.

ROBERT VAN GULICK

Qualia, Functional Equivalence, and Computation

Despite their impressive abilities to calculate and process information, present day computers do not have feelings, experiences, or inner lives involving qualia or phenomenal properties. Is this merely a reflection of the present limited state of computer technology or are there *a priori* and conceptual reasons which preclude the possibility of developing computers with qualia? If, in the future, robots are built which appear to display the full range of human affective behavior, how would we decide whether or not they did in fact have feelings and experiences? How could we determine whether they felt pains and enjoyed the taste of chocolate or merely simulated the human behaviors associated with such inner states?

These are the questions Moor addresses in his paper, "Testing Robots for Qualia." He hypothesizes the existence of a future robot that "behaves in a manner closely approximating human behavior and that ... has an internal organization which is functionally equivalent to relevant biological systems in a human being." (Moor: this volume, p. 108) He considers two sorts of tests which might be used to determine whether such a robot had experiences or qualia: the transmission test and the replacement test, each of which has a direct and an indirect version. He finds that none of these tests would decisively answer the robot qualia question. He concludes that the question is not subject to rigorous empirical test, at least insofar as one remains committed to the basically functionalist view of mind which he takes to be implicit in the tests discussed. He argues that attributions of qualia to robots would have to be largely non-evidential and would be justified instead on the basis of explanatory power with respect to robot behavior.

I am inclined to agree with Professor Moor about the inconclusive nature of the tests he considers, but to disagree about the general consequences for functionalism. The two sorts of tests he considers do not seem to exhaust the options open to the functionalist. Moreover his initial formulation of the problem threatens to make qualia epiphenomenal in a way which would undermine his proposal to justify qualia attributions on the basis of their explanatory power. Let us begin by considering his statement of the problem. Moor's robot is hypothesized to have an internal organization which is "functionally equivalent" or "functionally analogous" to the behavior controlling systems of the human brain. However it is not all clear what such equivalence requires.

At some points Moor seems to suggest it requires only input/output (I/O) equivalence; that is, the functionality of the system, the relationship of inputs and outputs, is maintained although the makeup of the system

H. R. Otto and J. A. Tuedio (eds.), Perspectives on Mind, 119–126.

is changed. I/O equivalence is a fairly weak relation and allows enormous variation in the causal system mediating inputs and outputs. It is not surprising that there might be systems which are I/O equivalent to humans but which lack qualitative or experiential states. Such a finding would have little impact on functionalism. I/O equivalence requires only simulation of human behavior, and most functionalists have denied that purely behavioral criteria can suffice for the application of mental predicates.

I suspect that Professor Moor has a stronger equivalence relation in mind since he writes of basing the robot's design "on the internal workings of a human being," and in his discussion of the replacement test he describes the substituted electronic component as functionally equivalent to the replaced brain portion. But he does not explain just what sort of equivalence this might be or in what respects the robot's internal workings (or Sally's electronic implant) are analogous to those in a human brain. This is unfortunate since the notion of functional equivalence is notoriously slippery [1] and *central* to the question at hand. Unless we can say what counts as *playing the same functional role* as a qualitative state (i.e. as being functionally equivalent to such a state), we can not hope to determine whether non-qualia states could play such roles.

Although he does not make an explicit statement on the issue, Moor seems to think of functional roles as nodes in a network of states defined by their relations to inputs, outputs, and one another, with the nodes linked by the relation of simple causation. That is, the state, behavior, or perceptual input at a node is linked to another if it typically causes or is caused by the latter. The network should also allow for causal inhibition and cases in which activation of more than a single node is required to produce a subsequent effect. Despite these complications the basic linking relation remains that of simple cause and effect.

This view is naturally associated with machine-state functionalism and with the popular technique of defining functional roles by the modified Ramsay method used by Lewis [2] and Block [3] insofar as the relevant network is interpreted only in input and output terms. The functional roles thus defined are quite abstract; the range of realizations is constrained primarily by the nature of inputs and outputs, and there is no procedure for requiring relations among nodes more specific than mere cause and effect. The method does not normally provide for more specific interactions such as requiring the occupant of node A2 to pass sodium ions or a certain string of binary code to the occupant of B3. In brief, it precludes requiring one node to bear any relation of qualitative or phenomenal similarity to another. Such models can only require that the occupant of A2 play *some* role in the causation, activation, or inhibition of B3.

If the functionalist is restricted to abstract causal networks of this sort interpreted only in terms of their perceptual inputs and behavioral outputs, it seems unlikely that he will be able to exclude non-qualia

realizations. But the moral to be drawn is not that functional descriptions cannot capture qualia, but rather that a richer vocabulary is required for specifying functional networks. Thus we have still not arrived at a satisfactory interpretation of our original question: could a robot which had an internal organization functionally equivalent to a human brain lack qualia? The notion of functional equivalence cannot be interpreted as I/O equivalence *or* as equivalence with respect to a simple cause and effect network of the kind just described without trivializing the question. On either reading the answer is probably, but uninterestingly, affirmative.

Delimiting the relevant notion of functional equivalence or functional role requires a principled way of distinguishing between *what some item does* and *how it does it*, which allows for the possibility that some structurally distinct item might do the same thing but in a different way. A fuse and a circuit breaker both prevent current from exceeding a certain maximum. One does so by melting as a result of heat qenerated by electrical resistance, the other opens because of electromagnetic repulsion. However, their description as functionally equivalent is principled only relative to a given level of abstraction and an associated context of pragmatic interests. If the context shifts to include other causal interests, such as interactions with nearby heat sensitive or magnetically sensitive components, the two will no longer count as functionally equivalent.

This well known relativity of functional equivalence [4] has important application to Moor's question. We want to determine whether a robot could lack qualia while having an internal organization functionally equivalent to a human brain in all *psychologically relevant respects*. We cannot require the robot's organization to be causally equivalent in *all respects*, for then nothing would suffice except giving the robot an artificial but molecule for molecule duplicate of a human brain. Any difference in composition which was perceptible or even indirectly detectable would constitute a difference in causal role. Thus, what is required is some relation weaker than total causal-role equivalence but stronger than I/O equivalence or simple causal network equivalence. We want a notion of *psychological equivalence*, but unfortunately that notion is itself far from clear. How do we draw the line between a brain component's psychological role and the non-psychological facts about how it fills that role? In fact, it seems there will be no unique way of drawing such a line; rather the line will shift depending upon our particular psychological inquiry.

In the case of a computer subsystem, we may be content to describe it in terms of its input/output function as a multiplier. But in other cases we may wish to push farther and distinguish between two such I/O equivalent units if one produces its results by serial additions and the other relies in part upon circuit analogs of multiplication tables. We will often wish to distinguish among devices that operate according to different algorithms, have different architectures, or employ different sorts of

representations, even if they produce similar outputs.

It seems likely that in at least some psychological cases, we will want to distinguish between systems with qualia and those without. Consider color qualia. They are most plausibly treated as properties of complex 3-dimensional representations. Normal visual perception produces representations with the formal structure of a 3-D manifold whose regions are differentiated at least in part by color qualia. Those colors also have a complex formal structure of similarities, unary/binary relations, and brightness relations. While it might be possible to process and store the information contained in the visual manifold in other non-qualitative ways, they would be importantly different from those involved in normal visual perception. Non-qualitative representations might be *informationally equivalent*, but they would have to be quite different in format, structure, and the nature of the processes which operated with respect to them.

Thus, if we are employing a notion of psychological equivalence which distinguishes among psychological subsystems on the basis of the sorts of representations and processes they employ, we will get a negative answer to our original question. No robot component could be functionally equivalent to such a brain system in the psychologically relevant sense unless it involved the use of qualitatively differentiated representations. The functionalist need not restrict himself to Professor Moor's two sets of tests. Rather, he can appeal to evidence about how the component subsystems of the robot operate. Just what sort of evidence he will need is at present uncertain, since we remain ignorant about the underlying physical basis of qualitatively differentiated representations in the brain. But we can reasonably hope that theoretical understanding of such matters will be forthcoming and may well arrive on the scene before the advent of convincingly humanoid robots of the sort Professor Moor hypothesizes.

Given an adequate theory of how qualitatively differentiated representations function and how such functions might be realized by underlying causal mechanisms, the functionalist would be prepared to address the robot-qualia question. There is no need for a transmission test. Neither direct nor indirect empathetic perception of robot qualia would be needed to establish their existence. Rather, it could be established in the standard scientific way by theory-based inferences from data about the robot's internal physical structure and activity, just as scientists today indirectly establish the existence of catalyzing enzymes in protein construction or photon-captures in photosynthesis. Scientific observation of qualia need not be empathetic.

Some versions of the transmission test could nonetheless be useful. If, for example, qualia should turn out to be associated with dynamic properties of electrical fields as certain Gestalt psychologists conjectured early in this century, a transmission might be devised which replicated in the perceiver the sort of fields occurring in the "brain" of

the subject being observed. The instrumentation for such a test would have to be based upon a prior theory about the underlying basis of qualia, but it would avoid the sorts of Lockean worries raised by Professor Moor. Given suitable theory and technology, empathetic perception might be possible to supplement indirect methods of non-empathetic observation.

A functionalist theory of qualia would also provide a more satisfactory formulation of the replacement test. The replacing component would have to do more than replicate the causal effects of Sally's damaged brain unit relative to verbal behavior, non-verbal behavior and the production of verbally encoded belief representations. It would have to have a physical organization of the sort needed to realize the functional properties theoretically associated with qualitatively differentiated representations. Without such a structure it might show the right sort of input/output activity, but it would not be producing those outputs *in the required way*.

Moreover, I am skeptical that non-qualia components could produce all the right outputs. As Moor notes, outputs include beliefs about qualia, and I am more sympathetic than he is to Shoemaker's claim that a creature without qualia could not have the relevant beliefs about qualia [5]. Though it is not quite Shoemaker's way of making the claim, a quick argument can be given to establish his point. One cannot believe a proposition one does not understand. A creature without qualia cannot fully understand what qualia are; so, such a creature cannot fully understand or believe propositions about qualia. Such a creature could not have beliefs equivalent in content to those which a normal human has when he believes that he is having a toothache, a red after-image, or is savoring the taste of a good Chardonnay. Thus, no component in a non-qualia robot could produce all the outputs produced by a normal qualia component in a human brain.

The functionalist equipped with an adequate theory would also be in a much better position to make the sorts of explanatory appeals to qualia that Professor Moor falls back upon at the end of his discussion. For without such a theory, it is not at all clear what explanatory work qualia are to do. In Professor Moor's hypothesized cases, the robot's internal workings are to be functionally equivalent to the behavior regulating portions of the human brain, while leaving the qualia question open. In such a case, what additional explanatory value could be purchased by attributing qualia to the robot? We might make the robot empathetically comprehensible to ourselves, but this would work as well for non-qualia robots as long as they simulated human behavior. Professor Moor does appeal to levels of explanation, and claims correctly that a complete description at the microphysical level will not suffice for every explanatory purpose. However, by allowing that any functional role filled by a qualia component might also be filled by non-qualia structures or processes, he deprives qualia of any causal explanatory role. He makes qualia (or at least the difference between qualia and non-qualia processes)

epiphenomenal. By contrast, the functionalist with a theory about how qualitatively differentiated representations function and are realized can invoke it to explain the causal operations of the relevant internal components, such as those underlying visual perception.

Professor Moor's distinction between qualia attributions based on *evidential* considerations and those based on *explanatory* considerations is not really viable. What we want is a theory which allows us to use detailed evidence about internal organization to explain how qualia function in the causation of behavior. Qualia attributions made in the context of such a theory would be genuinely explanatory. One final point requires mention. Moor sometimes asks whether we could build a computer with qualia and at other times whether we could build a robot or electronic device with qualia (the meat/metal distinction). Though he seems to regard these questions as interchangeable, they should be kept distinct. While most present day computers are electronic devices, not all electronic devices are computers. Nor need future computers be electronic. The *computational theory of mind* should not be confused with a commitment to *physicalism* or *mechanism*. Many critics of the computational view, such as John Searle, explicitly maintain a materialist view of mind [6]. The materialist is committed only to the claim that producing qualia requires building a system with the necessary physical organization. Computationalists claim that the relevant features of that organization are solely computational. According to them, having a mind, mental states, or perhaps even qualia requires only having a physical organization that instantiates an appropriate formally specifiable computational structure. No other physical constraints are placed on the class of systems with genuine minds. What physicalists like Searle object to is the suggestion that sufficient conditions for having a mind can be specified in such an abstract vocabulary unconstrained by any specific conditions on the details of physical constitution.

By analogy, we might apply the computationalist/physicalist distinction to the case of artificial genes. Is it possible to build robots or computers capable of "sexual" reproduction? The physical basis of human sexual reproduction and genetic transmission is today well established, and thus it is at least possible to construct artificial sexually reproducing physical devices. But it is far less obvious that doing so need only be a matter of making devices with an appropriate computational structure. Nor is it clear that electronic components could carry off the task. Consider how the claim Moor makes in his second to last paragraph about mimicking the computational structure of the brain would read if modified as a claim about genes: "Assuming human reproduction is directed by human genes and the genes operate through the activity of nucleotides which can be described by a computable function, a computer could be designed to instantiate the relevant part of the biological system." As an orderly physical process the activity of the nucleotides probably can be described by a

computable function, but not every realization of that function will be a system of sexual reproduction or an instantiation of the *relevant* parts of the original biological system.

The anti-computational physicalist claims that having thoughts, experiences, and qualia is more like being capable of sexual reproduction than like being an adding machine. Having the requisite sort of causal organization is not merely a matter of instantiating a certain sort of formal or abstract computational structure. More concrete causal constraints apply. Still, it may turn out that the relevant sorts of causal processes involved in having qualia can be produced in electronic as well as organic components. If, for example, the old Gestalt proposal identifying experience with electrical fields happens to be correct, such fields might be capable of production by non-organic components. But that result would still not confirm computationalist claims given our present ignorance about the physical basis of qualitative experience; for there is little we can say about the range of systems in which they might be produced. But in considering and investigating the question, we should be clear about the options and not confuse physicalism with computationalism, nor general questions about robots with more particular questions about computers.

END OF VAN GULICK'S PART 1

— ∇ —

EDITOR'S COMMENTS

Moor has assumed a position which aligns him with those functionalists who maintain that in order to understand the mind all that is needed is a *computational* replication of human skills and behavior. For them *any* physical organization capable of the requisite computational functions would suffice: "No other physical constraints are placed on the class of systems with genuine minds." (Van Gulick: this volume, p. 124) Van Gulick, however, notes that there is an important distinction which computationalists tend to overlook, namely that between *"what some item does* and *how it does it."* Given this distinction, the concept of "functional equivalence" used by Moor and the computationalists is inadequate to the task of describing relevant behaviors in contrasting systems. On the other hand, if an appropriate revision of this key notion is carried through, then their argument fails. And in that case the question about qualia reasserts itself.

Van Gulick argues, therefore, that the goal of functionalism should be to identify not only the computational structures that constitute the "behavioral organization" of the brain, but the *underlying causal mechanisms* as well. Whether these mechanisms are themselves ultimately reducible to computational description without remainder is an open question, one awaiting the judgment of further empirical research. If they do happen to turn out to be reducible, then if we could identify the physical processes which give rise to qualia in human beings, we could presumably generate a

computational translation that would replicate such processes within a system of programmed functions capable of running on "hardware" quite different from our own, and such a mechanism *would experience qualia*. The assumption here, of course, is that an "orderly physical process" can be replicated by a "computable function."

But how would we *know* that it experiences qualia? Would it be enough for us to focus on the "functionally analogous" behavior of the mechanism as Moor does? Van Gulick sees nothing to be gained from this line of questioning. He argues that we would have every reason to attribute the experience of qualia to an artificial intelligence whose structure was transparent to us at the design level and which exhibited all the proper behavioral patterns and responses. He asserts that,

> it could be established in the standard scientific way by theory-based inferences from data about the robot's internal physical structure and activity, just as scientists today indirectly establish the existence of catalyzing enzymes in protein construction .... Scientific observation of qualia need not be empathetic. (p. 122)

In other words, to the extent that qualia play a functional role in the internal organization of the mechanism's computational *and* underlying causal structure there is no reason that the experience of the mechanism would lack any of the qualitative dimensions intrinsic to human experience. Van Gulick would retain but reformulate the testing strategies discussed by Moor, for he feels they might be useful in helping us to comprehend the electrical dynamics of the brain system. But do qualia have the sort of nature that can be replicated in a functionalist format, with or without relationship to an underlying causality?

In the following commentary, Henry Johnstone proposes that qualia are mental phenomena that "emerge" from an *interpretive* process intimately bound up with communication. If qualia are indeed products of such "interpretive screening," functionalist replications of qualitative experience would need to include some sort of "translation" of the relevant interpretive functions. But, even given that, how would we tell whether the translation was a success? Johnstone's response would be to test robots for the ability to *use language*. Previously, Moor had argued that the qualia problem is analogous to the problem of other minds, and equally intractable. Johnstone's proposal appears to provide a countervailing view on this matter; for while Moor pushes the skeptical proposition that we cannot determine that qualia exist even in other minds (much less in robots), Johnstone, starting from the fact that we *do* attribute qualia to other minds, proposes that our use of language can be exploited to determine the presence or absence of qualia in human subjects. He concludes, consequently, that a "communication test" might work with robots as well.

HENRY W. JOHNSTONE, JR.

Animals, Qualia, and Robots

Moor's project is to suggest tests to determine whether machines experience qualia on the assumption that the analysis of the proposition that they do experience them is already settled--at least supposing that they are made of the right stuff. I assume this too. My project is also, in a sense, to suggest a test; but it is a test not so easily formulated as any of Moor's.

Moor asks whether the fact that robots are made of a stuff different from ours (metal, not meat) precludes their experiencing qualia. Is the stuff necessary for the experience? An equally legitimate question is whether the stuff is sufficient for the experience, assuming a properly functioning organism. Do animals experience qualia? We have little difficulty in supposing that monkeys, dogs, and cats do. What of lizards? Bees, we know, are sensitive to many wavelengths in the light spectrum, including wavelengths we are not sensitive to; but do they experience *the qualia* of red, blue, and ultra-violet? And what would be added to our understanding of bee behavior by saying that they do? With bees, we seem to be confronted with machine-like objects, and Moor ought to find it just as plausible to suppose that bees have qualia as that machines do--and for the same reasons. If transmission and replacement tests can (at least in principle) be designed for machines, there is no reason why they cannot (at least in principle) be designed for bees. Descartes held that animals *are* machines. One thing that prevents many people from accepting this contention is their inclination to attribute qualia to some animals; e.g., cats, dogs, and monkeys. But with respect to bees, there is not a strong inclination to do this, and hence no great reluctance to agree with Descartes.

If it is reasonable to generalize the qualia question from machines to animals, the question can be further generalized in an interesting way. For the question whether non-humans experience qualia is analogous to the question whether and in what sense non-humans use language. Bees are again a good example. There is clearly a sense in which bees use language. Their dances communicate the whereabouts of nectar. The sense of "communicate" here is the same as that in which machines "communicate" with one another. A radar beacon can communicate to the computer of an airplane the whereabouts of an airport. Is there any need to assume that the bees are communicating with one another in a stronger sense of "communicate" than that in which it is sufficient to assume that the machines are communicating? This is like the question "Is there any need to assume that bees are sensitive to colors in a sense of 'sensitive' stronger than that

H. R. Otto and J. A. Tuedio (eds.), Perspectives on Mind, 127–136.

in which it is sufficient to assume that some machines are sensitive; i.e., that they are responsive to stimuli without experiencing qualia?"

It would be helpful if one could offer a clear definition of this stronger sense of "communicate," because we might then see what is meant by the stronger sense of "sensitive," but the task is not easy. Perhaps we could start by claiming that an airplane pilot in contact with a ground controller is communicating or being communicated to in the stronger sense, because as the result of what the controller tells him, the pilot *knows* where he is. But how will the knowing pilot differ from the autopilot? Unless he operates the controls of the plane in exactly the same way (or very much the same way) that the autopilot would have operated them, he cannot be said really to know at all. Perhaps the pilot "knows" in the sense that he could give an account of the plane's whereabouts to another person. But the onboard computer, in collusion with the beacon, and on the basis of other signals, could probably give a better account. We are on a slippery slope. As long as we treat communication, or the understanding of what is communicated, as a competence, we can always design machines more competent than humans in exercising the skills that we claim humans exercise which they communicate in a sense of "communicate" not applicable to bees. This is an instance of the principle that however we define intelligent behavior, someone can design a machine capable of such behavior. Hence there is no difficulty in showing that machines are intelligent. We capitulate to this conclusion by failing to take issue with the assumption that intelligence is a form of behavior. What has gone wrong similarly with our attempt to isolate a kind of communication higher than that of the bees is that we have failed to resist the assumption that communication is or results in a sort of competence.

We come somewhat closer to grasping the sense of "communication" we are seeking if we see it as having a rhetorical dimension. Rhetoric is required at least to *call attention to* the content communicated. I must get the attention of my interlocutor or audience if I am to communicate anything at all to him or it. I must bring it about that minds are put onto what I am saying. A pilot *knows* where he is only if his mind is on his whereabouts. He can, or course, respond to signals like an autopilot, but I think we would characterize such response as automatic or purely reflexive, like the response of the bees, not based on knowledge. The dancing bees engage in no rhetoric of attention-getting, nor does the radar beacon addressing the autopilot. No such rhetoric is necessary, since the members of the dancers' audience have no choice except to respond, and the same for the autopilot. But the attention of the pilot must somehow be drawn to his own situation if he is to be said to "know" where he is. It does not much matter what alerts it--whether another human in the cockpit, or the pilot somehow collecting himself together or simply spontaneously noticing some signal or reading showing on an instrument. My point is not

that machines are never the *source* of the rhetoric of attention-getting; it is that they cannot be its *destination*. A person *knows* his whereabouts only when his *attention* to his whereabouts has been summoned by some rhetorical stimulus. But it makes no sense to speak of getting a machine's attention; once it is properly switched on and tuned in, it has no choice except to lend its ear.

There is another way to put the point. The relation between the dancing bees and their audience is a dyadic relation; the dancers stimulate the foragers, and the latter respond. But a dyadic relation is an inadequate model of communication in any except the rudimentary and perhaps metaphorical sense in which the bees are said to communicate or a radar beacon may be said to communicate with an autopilot. As Peirce plainly says, three terms are needed: a sign, that of which it is a sign, and the being that interprets the sign. Thus the ground controller's words are a sign *to* the pilot *of* the whereabouts of the airport. But we cannot formulate a corresponding analysis of the "communication" of the bees. For it would sound very strange to say that the dance is a sign *to* the bees *of* the whereabouts of nectar. That would suggest that the bees had their minds on a task. And such a suggestion contradicts our understanding of insect behavior as mindless--as based on unconditioned reflex rather than intellect.

My appeal to the concept of rhetoric in attempting to characterize communication at a level above the most primitive is not, however, equivalent with my appeal to Peirce's semiotic triad for the same purpose. For we can catch the attention of a being not certainly capable of taking the role of interpretant. A playful or distracted dog may have to be addressed repeatedly before it will finally listen to a command. But once we have gotten its attention, can we be sure that it will interpret the command as a sign of something else? This seems unlikely, for what we mean by a "command," at least as addressed to an animal, is a stimulus intended to elicit a response. Rhetoric, in other words, may be a prerequisite to semiosis, but can also be a prelude to communication at a pre-semiotic level, at least with some animals.

Communication with dogs is, of course, a two-way street, and when the other interlocutor is a human being it can be genuinely semiotic. *To* a human the dog's bark can be a sign *of* someone's being at the front door. But of course exactly the same can be true when the source of the sign is an inanimate object such as an instrument in an airplane. And in both cases communication reaches the semiotic level only because of the human participant in the interaction.

What if the dog could understand its own bark as a sign of the presence of someone at the door? Then it would be telling us something in a way most people would regard as uncanny. The dog would itself be communicating at a level above the rudimentary and metaphorical, because

its mind would be on the task of getting a message across.

This possibility is not often seriously discussed in connection with dogs, but the question has, in effect, been raised whether *chimpanzees* can be interpretants of their own signing behavior. If they can, the question whether they are language-users can be answered in the affirmative. If not, this behavior, however complex it may be syntactically and in vocabulary, seems to reduce to a tactic of problem-solving. (The dog is presumably also trying to solve a problem by barking.)

Is there a language in which humans and chimpanzees can communicate? So far as the human participants are concerned, there clearly is. These humans are aware of the semiotic function of their own messages to the chimpanzees and respond to messages from the latter as signs, not just stimuli. But it can still be a moot question whether the chimpanzees see their own messages as signs, or, for that matter, whether they are interpretants of the messages they receive from humans. How would one find out? Any objective test of their status as interpretants could in principle amount to no more than a test of their responses to certain stimuli, and thus would be self-defeating. The only hope would be in *asking* the chimpanzees whether they are interpretants. We would have to pose questions like "Does the yellow disc mean 'banana'?" Such questions would clearly require a much richer vocabulary and syntax than any hitherto taught chimpanzees. In order to be reasonably sure that the response was not merely the result of training, we would have to insist on the use of a language comprehensive enough to allow question and answer to be framed in a number of ways, as well as to permit excursions into the metalanguage (as in "X means 'X'.") The result would be that any chimpanzee able to understand the question would automatically qualify as an interpretant.

I return to the question of qualia. Qualia emerge in an interpretative process. When this process is absent, objects come close to being pure stimuli. Thus the ring of the telephone can stimulate a reflex arc causing me to pick up the receiver. In this case I hardly notice the sound at all. I do notice it when I am further away from its source, and the ring contrasts with other background noises; then the ring emerges as a quale. But I can fail to notice this contrast, too, if my mind is not on what I am hearing; there may be no qualia for me at all if I am daydreaming.

The interpretative process in which qualia emerge is a kind of discrimination. But "discrimination" is an ambiguous word; it can be a response to pure stimuli, as it probably is in the case of the bees who fly to the sugar-water in the red dish but not in the blue. This feat can be explained without any reference to an act of interpretation. But if humans choose the contents of the red dish over those of the blue, we do not assume that they are conditioned by a stimulus that does not enter into their experience. It is more plausible to suppose that the very

distinction between red and blue arises as a way of marking the distinction between positively valued contents and contents without this value; that this distinction is called for as an interpretation of the *value* distinction. If there were no value distinction, the color distinction would no longer matter, and would tend to fall from view, to lapse.

It would be preposterous to deny that humans can entertain qualia wholly apart from their role as markers of distinctions. It would similarly be preposterous to deny that a sign can be enjoyed for its own sake, wholly apart from the object of which it is a sign. Painting and poetry--not to mention all the other arts, or indeed whatever induces esthesis--would be devastating counterexamples to any such thoughtless denials. It can nonetheless be reasonably claimed that the qualia entertained in esthetic experiences must first be gained as vehicles of distinction. This process can be documented in art itself, which invites us not only to entertain but also to discriminate.

Just as signs can be characterized in terms of their position in a triad, so can qualia. What distinguishes a quale from a pure stimulus is that *someone* ascribes it to *something*; e.g., I ascribe the red to the dish. The similarity between this process and that of semiosis is obvious. Qualia are, in fact, signs.

When Moor speaks of qualia, his examples are not primarily colors or sounds. They are pains. The crucial problem for him is whether a robot can experience pains as Sally does. But there seems to be no problem about fitting pains into the triad. A pain is distinct from whatever causes a reflex flinching; that would be a term in a dyadic relation in which pain had not yet arisen. Pain does arise when *someone* ascribes a certain quale to *something* (using "someone" in a broad enough sense not to rule out animals or machines). For example, I ascribe a pain to my tooth. The word "ascribe" here may seem a little strange, since I probably have no *choice* except to feel pain in my tooth. The act of ascription, in the sense intended here, is not the result of reflection. If there is a more suitable word than "ascribe," let it be used.

While it is likely that I have no choice except to feel pain in my tooth, I can--not through choice--fail altogether to notice it, as when my mind is on something else. Similarly, a person for whom a message is intended can, through absent-mindedness, fail to play the role of interpretant. For semiosis to occur, there must be attention. Similarly for qualia.

How do we know whether an organism or machine experiences qualia? Objective tests can in the end do no more than provide stimuli for the subject to respond to. If I flash a red light and you say "red," how do I know that you are experiencing the quale red? And if this is the case for people, it is *a fortiori* the case for animals and machines. The only hope will have to be what was the only hope of learning whether a subject is an

interpretant--namely, the use of questions and answers in a language sufficiently comprehensive and open to allow us to discuss the subject's experiences with him. I assume, incidentally, that the possibility of such discussions in the case of humans is a powerful argument against solipsism, which in the face of a flexible and reflexive language becomes a hopelessly complicated hypothesis.

If we learn through conversation whether a subject experiences other qualia, we learn about the subject's pains in this way, too. So the question is whether animals and machines can use a suitably complex language to enable them to discuss the matter with us. This conclusion reverses our deepest intuitions about the capacity for pain of animals and that of machines. We think it probable that at least the higher animals experience pain and improbable that machines do. And yet it is far more likely that machines can be constructed capable of using languages of the requisite complexity than that any animals exist with such a capacity.

But perhaps there is a confusion here. Animals exhibit "pain behavior"; they flinch and scream when confronted with certain stimuli. Machines do not usually behave in this way. But the behavior in question has little bearing if any on the issue of qualia. That issue is, I take it, epistemological; we want to know what data can be available to subjects of various sorts. A quale is a datum; a flinch is not. Or, as Moor puts it, a quale presupposes a belief. But only a linguistically sophisticated being can formulate its own beliefs. Again, to refer to another of Moor's examples, the issue is not just pains but the *identity* of pains, and such identity is neither asserted nor established by a scream.

Quite distinct from epistemological questions, and of far greater practical importance, is the moral question of how to deal with animals and robots. It is obviously cruel gratuitously to stimulate pain behavior, especially on the part of organisms unable to formulate their reactions to the stimuli--able only to scream. A scream is not a report, but it can be a powerful moral imperative.

—— ▽ ——

Johnstone takes issue with Moor's unspoken assumption that intelligence is a form of behavior involving the exercise of skills and competences that can be replicated in a computational mechanism. If we were to ask Moor how he would determine when a robot is exhibiting intelligence, he would respond by analyzing the behavioral traits of the robot. Similarly, to determine whether or not the robot is *communicating* with us, he would look for language performance that is functionally analogous to our own. Johnstone is critical of this approach primarily because it overlooks a crucial feature of genuinely intelligent behavior.

This feature, which sustains our capacity to use language as well as to experience qualia, is explained by Johnstone in terms of the role played by "attention" in shaping our linguistic response to stimuli.

Qualia are said to differ from mere stimuli. A bright light will set in motion a "reflex arc" behavioral response: we blink, or we turn our head away from the light. But what about the painful glare of that light even as the reflex swiftly completes itself? Or, what about a noise in the woods that makes one's skin crawl at night? Why do some sounds catch my attention, while others do not? What about my preference for dark blue over light blue? At what point are we beyond examples of mere stimuli? What is it that gives rise to *qualia*?

Johnstone argues that the key to experiencing qualia lies *not* in exhibiting the proper behavior, but in having the internal capacity to identify stimuli as meaningful signs. The ability to discern qualia depends on our capacity to identify, from a first person standpoint, the special characteristics of stimuli infused with significance or meaning relating to *our* situation. This, in turn, implicates the presence of sophisticated linguistic capacities allowing the creature in question to formulate beliefs in ways that can be communicated to others. Ultimately, then, the existence of qualia appear to go hand-in-hand with the ability to express meaning through channels of communication.

Moor's analysis of the qualia issue has stimulated reflection on some important issues; however, major questions remain unanswered. Since the testing problem (with respect to the metaphysical side of the qualia issue) appears not to have been defused, we may yet have to deal with the issue of how to test for qualia in mechanisms designed to operate as "full-fledged" minds. Nor have we anything more than a provisional understanding of the impact of qualia on behavior. It has been argued that the "taking" function plays a key role in the experience of qualia, but not enough has been said about the nature of this function to determine whether or not it lends itself to computational reduction. We need to examine the possibility that "taking" is an intrinsic, intentional structure of mental processing. If we can determine that intentionality *is* integral to the "taking" function, then we are brought once again to the central question raised in Chapter One, namely, to what extent is the *semantic* content of mental processing reducible to the formal syntax of computable functions?

Computationalists, of course, face an additional challenge from those who argue that *physical organization* plays a key role in the structural makeup of qualitative experience. Proponents of this view contend that, sooner or later, the computational theory of mind must wrestle with the problem of replicating the causal organization of neurobiological functions. Here the computational strategy will meet its match, these critics argue, for the requisite physical-chemical organization cannot be translated into the formal syntax of computable functions. Lacking the

ability tonalize this crucial element of the puzzle, computationalism is stoppeort in its attempt to design a network of computational functionsble of replicating a full-fledged mind.

We egin tackling these issues in the next section which revolves around ar by R.J. Nelson. In the context of his investigation of the "taking" tion, Professor Nelson proposes an eclectic merger of physicalisd computationalism designed to free functionalism from the constraint would otherwise be imposed upon it by a purely computational aph. Were he to succeed, he would thereby fulfill the requirement set by Van Gulick in the latter's critique of Moor. But, in the end, he irs to be driven to admit that even this eclectic merger may fall short account of "full blown" intentionality.

2.2 Inteality

The ntional character of mental activity is a primary object of reflection most "first person" approaches to the study of mind. In contrast, it (if not all) computational and physicalist approaches continue tumble against the enigma of intentionality. Indeed, as we witnesse the Rey and Moor essays, there is a growing tendency from these sta nts to ignore or downplay the importance of intentionality as a charact of mental life.

Thewing paper, by R.J. Nelson, attempts to take the middle road: while acng that the intentional character of mental life is an importanredient that must be accounted for by an adequate theory of mind, N tries to show how we might analyze intentionality from a standpoiat merges the computational and physicalist approaches. He calls his oach "mechanism," and contends that while it may fail to capture ssence of conscious intentional attitudes, it is nevertheless sufficien analyzing all other forms of intentional phenomena, including full range of our perceptual experiences and our "tacit" beliefs asires.

Nel tresses the importance of viewing the mind holistically, but argues many of the holistic aspects of mental life can in fact be analyzed rms of computational and/or neuro-biological functions. The early sec of his paper present a criticism of the purely computational approach the study of mind, highlighting weaknesses inherent in the function conception of feelings, beliefs and desires as "role-playing" tal states. He also stresses the importance of distinguishing "mental udes" (which are intentional in character) from "cognitive skills reig intelligence" (which need not be intentional). Thus he proposes:

> a computer might "read" stereotypical print, "play" chess, "compose" music, "draw" pictures, and "prove" theorems, but would not believe, perceive, strive for, hope for, or understand a thing. (Nelson: this volume, p. 145)

Of course, the computer might prove to be functionally identical to a conscious entity that really does believe, perceive, strive for, hope for, and understand things. But this would demonstrate nothing more than token-token identity between the two structural systems. Lacking the neural network that instantiates our intentional life, and despite the programmed presence of token-identical logical structures, the computer would be incapable of feelings, sensations, or other subjective experiences, and hence would lack "full-blown" intentionality.

In contrast to Moor, Nelson proposes that the intentional life of mind is dependent on the right (neurological) *stuff*, and that every conscious mental occurrence is thus type-type identical to an event in the nervous system. This in turn leads Nelson to emphasize a sharp distinction between components and structures: "*components* are type-type identical to material events, *structures* are individuated functionally and are token-token identical to material complexes." (p. 147) This is the key to his mechanist version of computationalism. From this standpoint, he proceeds to analyze several key aspects of the intentionality issue, including the "taking" relation (which, given its semiotic character, necessitates an account of the self-referencing character of "recognition states").

As we saw earlier, Johnstone's reflections on the semiotic character of the "taking" relation foreshadowed Nelson's proposals. Johnstone emphasized the interpretive process that underlies all qualitative experience and focussed on the semiotic character of the relation that holds between "sign," "referent," and "interpreter." He contrasted this with a merely "dyadic" relation which structures the operations of computational mechanisms, and concluded that it is our capacity to focus "attention" in selective ways that separates us from these mechanisms, and separates us in ways that cannot be captured by computational-based philosophies of mind. He concluded that while the computational mechanism might be capable of manifesting behavior-tendencies that simulate our own, and might even manifest linguistic capacities good enough to pass a communication test, such a mechanism could not function in ways that are dependent on the "taking" relation, and so would lack the capacity for experiencing qualia or other intentional phenomena.

But what is this "taking relation?" In particular, is it really dependent (as Johnstone's discussion of "attention" has suggested) on *conscious* mental processing? Nelson will propose, contrary to Johnstone, that there is no reason why computational mechanisms could not be programmed to exhibit functions dependent on the taking relation. Here the