THE BEHAVIORAL AND BRAIN SCIENCES (1978), 4, 515–526
Printed in the United States of America

Does the chimpanzee have a theory of mind?

David Premack
Department of Psychology,
University of Pennsylvania,
Philadelphia, Penna. 19104

Guy Woodruff
University of Pennsylvania Primate Facility,
Honey Brook, Penna. 19344

Abstract: An individual has a theory of mind if he imputes mental states to himself and others. A system of inferences of this kind is properly viewed as a theory because such states are not directly observable, and the system can be used to make predictions about the behavior of others. As to the mental states the chimpanzee may infer, consider those inferred by our own species, for example, *purpose* or *intention,* as well as *knowledge, belief, thinking, doubt, guessing, pretending, liking,* and so forth. To determine whether or not the chimpanzee infers states of this kind, we showed an adult chimpanzee a series of videotaped scenes of a human actor struggling with a variety of problems. Some problems were simple, involving inaccessible food – bananas vertically or horizontally out of reach, behind a box, and so forth – as in the original Köhler problems; others were more complex, involving an actor unable to extricate himself from a locked cage, shivering because of a malfunctioning heater, or unable to play a phonograph because it was unplugged. With each videotape the chimpanzee was given several photographs, one a solution to the problem, such as a stick for the inaccessible bananas, a key for the locked up actor, a lit wick for the malfunctioning heater. The chimpanzee's consistent choice of the correct photographs can be understood by assuming that the animal recognized the videotape as representing a problem, understood the actor's purpose, and chose alternatives compatible with that purpose.

Keywords: chimpanzee communication; cognition; consciousness; intentionality; language; mind; primate intelligence.

Fifty years ago Köhler (1925) carried out his now widely known studies demonstrating problem solving and simple tool use in the chimpanzee. He confronted his chimpanzees with food that was inaccessible in a variety of ways – out of reach overhead, outside the cage mesh, locked in a box, and so forth – and found that in nearly all cases the animals would use existing cage materials to obtain the food. In other studies, Köhler set fruit-laden baskets into motion and then observed the animals leap to a ledge on the wall and arrive just in time to be met by the basket, suggesting that they could extrapolate future positions of the basket from current ones. Indeed, the animal may reveal his comprehension of physical relations even as he casually diverts himself. When the chimpanzee Elizabeth (Premack, 1976) sat in the hall waiting for her "language" lesson, she often recovered a ball that rolled away from her, not by reaching for it directly, but by pulling in the blanket on which it had come to rest. She pulled with the "right" force: the ball did not roll off the blanket but came steadily toward her. The chimpanzee's evident comprehension of physical relations makes it of interest to determine at what level these relations are available to the animal and whether there is a sense in which he can be regarded as a lay physicist. But questions of this kind are only indirectly relevant to our present concerns. We are less interested in the ape as a physicist than as a psychologist (every layman, of course, is both); we are interested in what he knows about the physical world only insofar as this affects what he knows about what someone else knows.

In this paper we speculate about the possibility that the chimpanzee may have a "theory of mind," one not markedly different from our own. In saying that an individual has a theory of mind, we mean that the individual imputes mental states to himself and to others (either to conspecifics or to other species as well). A system of inferences of this kind is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior of other organisms. We will not be concerned at this time with whether the chimpanzee's theory is a good or complete one, whether he infers every mental state we infer and does so accurately, that is, makes the inferences on exactly the same occasions we do. These questions are not out of order, but they are, for the moment at least, too difficult to deal with experimentally. It will be sufficient now to consider whether or not the chimpanzee imputes mental states to others at all. If we succeed with that claim, we may later seek to determine how accurate and complete his inferences are.

As to the states of mind the chimpanzee may infer, let us look at some of those that members of our own species infer. It seems beyond question that *purpose* or *intention* is the state we impute most widely; several other states are not far behind, however. They include all those designated by the italicized term in each of the following statements: John *believes* in ghosts; he *thinks* he has a fair chance of winning; Paul *knows* that I don't *like* roses; she is *guessing* when she says that; I *doubt* that Mary will come; Bill is only *pretending*. This list is in no way exhaustive. *Promise* and *trust,* for example, are important states not on the list, and there are other, more exotic ones, belonging to the novelist; we will not be concerned with them in this paper. We will also leave aside for the moment the embedded inference, for example, "Mary knows that John thinks he will win," "Harry doubts that Mary knows that John thinks he will win." A great deal of complexity can be imparted, if even only a small inventory of different states is linked together in this fashion. Human limits on

© 1978 Cambridge University Press 0140-525X/78/DPRE/011$04.00

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

embedding are not impressive: only about four steps make our species uncomfortable. And, of course, it would be a surprise if the ape could approximate even this limit. In the rest of this paper we will describe the experiments that defend the present thesis (more in principle than in fact, for as yet only a few of them have been done), discuss alternative positions, showing how these can be countered either by existing or potential evidence, and conclude by examining the implications of the present view for theories of intelligence and mind.

Basic approach: problem comprehension, not problem-solving performance

Rather than confronting a chimpanzee with an inaccessible object and observe his possible problem solving, we have instead shown him a human actor confronting inaccessible objects, and asked the animal to indicate how he thought the human actor would solve his problem. Specifically, we made four thirty-second videotapes of a human actor in a cage similar to the chimpanzee's struggling to obtain bananas that were inaccessible in one of four different ways: They were (1) attached to the ceiling and thus out of reach overhead, or (2) outside the cage wall and thus horizontally out of reach, (3) outside the cage, but with the actor's reach impeded by a box inside the cage, located between him and the bananas; in the last case (4), not only was the actor's reach impeded by a box, but the latter was laden with heavy cement blocks.

In addition to the four videotapes, we took still photographs of the human actor engaged in the behaviors that constituted solutions to the four problems. In one case, he was photographed stepping onto a box; in a second case, [lying on his side and reaching out of the cage with a rod;][a] in a third case, moving a box to the side, and in the last case, removing cement blocks from a box.

A comprehension test for the animal (Sarah) consisted of showing her each of the videotapes in turn, putting the last five seconds of the tape on hold, and then offering her a pair of the photographs, one constituting a solution to the problem and the other not.

Sarah, a fourteen-year-old African-born chimpanzee, was less than a year old when received in the laboratory and has since been trained and tested in a number of ways (Premack, 1976). When she was between 4.6 and 6.5 years of age, she was taught a simplified visual language. In addition, she has been tested on a variety of cognitive tasks: reconstruction of disassembled objects, causal inference, and so forth, five days a week for the last ten years. She had had no formal experience with the present tests, although she had had extensive prior experience with commercial television, which doubtless contributed to her ability to comprehend televised representations.

In the tests with the present material, Sarah was given each videotape on [four][b] occasions, each time with a [different][c] pair of alternatives. Each alternative was used equally often with each other one, and, of course, the left-right position of the correct alternatives was counterbalanced over problems and over trials. The general manner of testing, described in detail elsewhere (Premack & Woodruff, 1978), was designed to control for inadvertent social cues. Briefly, the trainer showed Sarah the videotape and then put the last five seconds on hold. He then presented two alternatives in a cardboard box and left the room. Sarah's task was to take out the alternatives, choose one, and identify her selection by placing it in a designated location alongside the television set; she was then to ring a bell, which summoned the trainer, who re-entered the room, recorded Sarah's choice, and told her either, "Good Sarah, that's right," or "No, Sarah, that's wrong" in a tone of voice like the one we would use with a young child. At the end of each session, he gave her yogurt, fruit, or some other favorite food. She was correct on twenty-one out of twenty-four trials ($p < .001$), all of her errors confined to one problem: removing blocks from the block-laden box.

Probably it is not a coincidence that this is the problem with which Köhler's apes had the greatest difficulty. The strength of the chimpanzee is such that she is not likely to appreciate the advantage (or the necessity) of removing the heavy cement blocks before attempting to move the box. In any case, this is the problem in which Köhler's apes largely failed, and the only problem with which Sarah had any difficulty. She made three errors in a row on this problem, but succeeded on the last three trials. She could easily have learned the correct solution in the course of being tested, but this would be an unlikely interpretation of her overall performance. She chose correctly from the beginning on the other three problems, and was correct six out of six times on each ($p = .013$ binomial test).

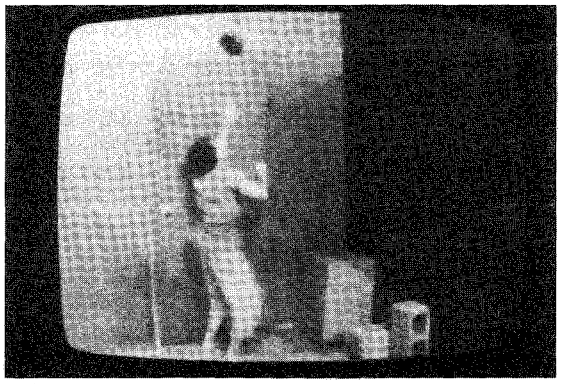
Three interpretations of chimpanzee comprehension of problem solving

Now consider three interpretations of this outcome, the first of which is easily eliminated. On this first interpretation, the animal simply matches physical elements in the correct alternative to the corresponding physical elements in the videotape. The possible elements here are the human actor, bananas, cement blocks, sticks, a rope, and a box. It was possible to rule out this alternative quite directly by arranging to include every one of the elements involved in every videotape. Consider Figure 1, which shows in the left column the last scene in each videotape, and in the corresponding right column the correct solution to the problem. It can be seen that every element is present in every scene. Figure 1 also shows that the posture of the human actor could have contributed to the definition of the problems. The actor was decidedly more upright in one case than in the three others (both in the videotape defining the problem and in the photographic alternatives constituting solutions). Yet matching the actor's posture in the problem to his posture in the solution cannot account for the animal's performance. She was correct on three of the four problems, whereas "upright" versus "nonupright" will only distinguish one problem from the other three. Moreover, the possibility of physical matching is ruled out not only by this series, but even more so by others we will report later. In general, physical matching is too weak a device to account for the results. Not only chimpanzees, but probably lesser species as well, can solve problems with strategies more sophisticated than simply matching physically identical or similar items.

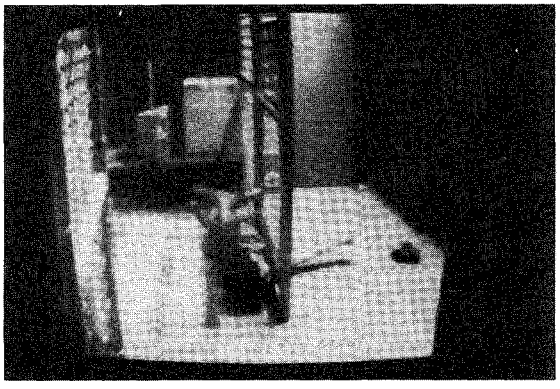
We turn now to two families of alternatives more in keeping with the power or complexity of the primate mind, and between which it will be necessary to decide.

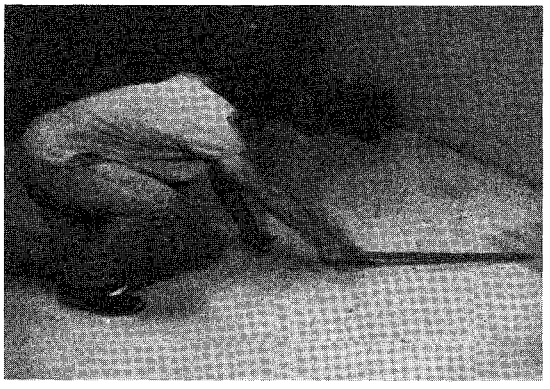
Associationism. The first of these is some version of classical associationism. One solves problems of the present kind, on this view, through familiarity with the sequences in question. When shown a sequence that one recognizes, but that is incomplete, one chooses the element that has the effect of completing the sequence. An even simpler version of this general view might be derived from a theory of interrupted action. If an animal is interrupted when carrying out an act, he will recommence and complete the act as soon as possible. An animal *shown* an interrupted action will do the same, as long as he is intelligent enough to recognize representations of actions. Species such as chimpanzees will thus do at a representational level what lesser species can only do at the level of direct action. But in both cases nothing more is involved than producing the next step in a known sequence. To apply this theory to the present results, we need only assume that the animal is familiar with the problems shown in the videotapes. And this is not an unreasonable assumption. Even though the animal has never seen either a human subject or a conspecific facing problems of the kind

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

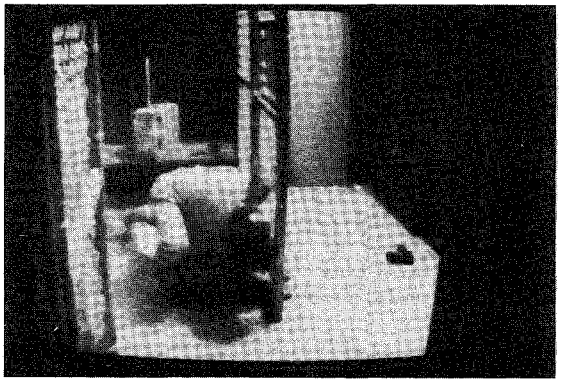


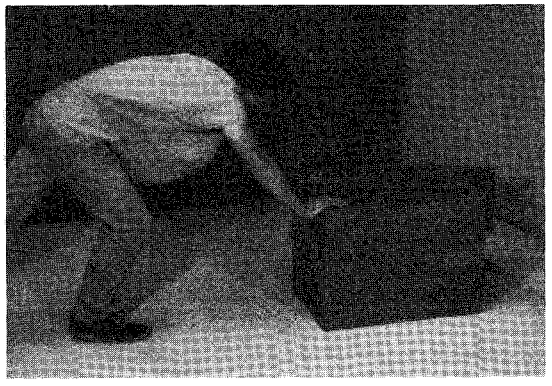
1



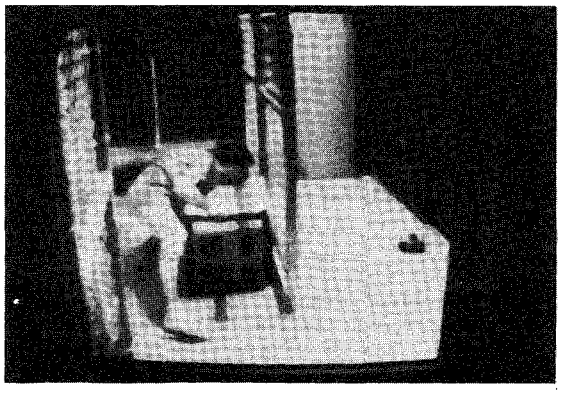


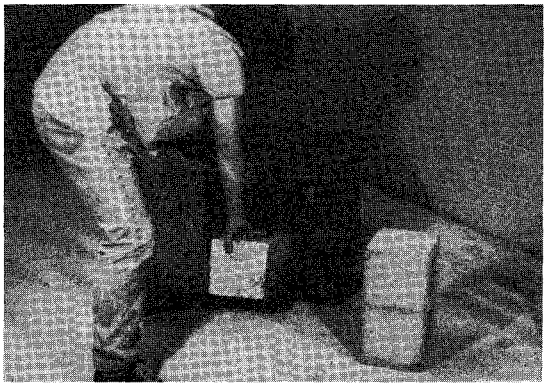
2





3





4

Figure 1. Photographic reproductions from the last 5 sec. of each of the four 30-sec videotaped problem scenes in Test 1 appear in the left column. In the right column are shown the still photographs of the correct solutions. For example, in Problem 1, the human actor attempts to reach up toward bananas suspended by a rope from the ceiling, and in Solution 1, the human actor steps on a box. See text for fuller description.

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

depicted in the videotape, she has herself encountered versions of similar problems informally in her daily life. It is hardly possible to prevent a chimpanzee from engaging in problem solving. As Köhler remarked, the first official problem solving his apes performed was certainly not the first in which they had ever engaged. In the course of her daily life, Sarah has had occasion to reach out for inaccessible objects and to climb onto structures in her cage in order to reach objects on the ceiling. The items she used were never boxes or sticks, nor were the objects she sought bananas. And of course it was Sarah who sought the objects, not the trainer whom she observed in the videotapes. Nevertheless, there was probably enough similarity between the form of her direct experience and the form of what she was shown in the videotape to grant the associationist argument. She was familiar with at least some of the sequences involved, so that when shown an incomplete version of them, she was capable of selecting the alternatives that completed the sequences.

The weakness of the associationist view has always been its difficulty in dealing with the future. How does the animal predict the next event in sequences that are not familiar? The associationist reply is well known: Success comes from the similarity between old and new cases, a similarity that allows generalization to do its job. The generalization proposal is appealing – but only so long as the data are vague and the sequences in question are ill defined. Under such conditions one has no alternative but to take the generalization proposal seriously. However, once the sequences are as well defined as they are, for example, in language, it is immediately evident that generalization is a vacuous proposal. No theory of generalization will explain the comprehension and production of structurally novel sentences. Instead, in cases of this degree of clarity we see not only the vacuity of the generalization proposal, but also, and more important, the form of the kind of theory that is needed. *Rules* that have the power to *generate* the well-defined sequences in question, not generalization gradients, are what would account for the productivity.

In the present article, however, we are not in the admirable position of the linguist. We will be concerned with many varieties of problem solving, for most of whose sequences a well-formed structural description is not yet possible. Consequently, we cannot convincingly rule out generalization: there exists just enough vagueness to allow the proposal its appeal. That is, we cannot rule out generalization on the basis of well-defined sequences; but there are other, equally compelling grounds on which we can eliminate an associationist interpretation, and when we reach that stage in the argument, we will point them out.

Theory of mind. The view we recommend, and will attempt to support here, is that the chimpanzee solves problems such as the present one (and others a good deal more complicated) by imputing states of mind to the human actor. In looking at the videotape, he imputes at least two states of mind to the human actor, namely, intention or purpose on the one hand, and knowledge or belief on the other. The chimpanzee makes sense of what he sees by assuming that the human actor wants the banana and is struggling to reach it. He further assumes that the actor knows how to attain the banana, so that when the animal is shown photographs depicting solutions to the problem, he chooses correctly in three out of four cases.

Empathy. In addition to the two major alternatives, there is the empathy alternative. The animal sees the human actor struggling to reach the bananas, "puts himself in the place of the actor," and chooses an alternative in keeping with what he would do were he in the actor's predicament. The animal's choice, on this view, is not a prediction of what the actor will do, based on inferences that the animal makes about what the actor knows, but is simply a prediction of what he would do were he in the actor's situation. It is clear enough how to distinguish the two interpretations. The empathy view will be supported if the identity of the actor has no effect on the animal's choices. Conversely, the empathy view cannot be correct if the actor's identity does affect the animal's choices. We will return to this point. In the meantime, it is important to note that empathy and "theory of mind" are not radically different views; they are in part identical. The empathy view starts by assuming that the animal imputes a purpose to the human actor, indeed understands the actor's predicament by imputing a purpose to him. The animal takes over the actor's purpose, as it were, and makes a choice in keeping with that assumed purpose. The empathy view diverges only in that it does not grant the animal any inferences about another's knowledge; it is a theory of mind restricted to purpose. It might be called a theory of mind concerning the other's motivation, as opposed to a more nearly complete theory that takes into account not only the other's motivation, but his cognition as well.[1]

Ultimately, however, we should avoid the impression that associationism and theory of mind are mutually exclusive. We have no objections to the former, except inasmuch as it is proposed as a sufficient and exclusive mechanism. Certainly, in highly familiar situations, one's expectancies are based on existing associations; they are not rule-generated. In novel situations, however, one's expectancies are generated, we think, from theories, and are not the product of associative generalization. This seems clearly the case in language, and is so, we believe, in other realms; language is not the only realm in which one builds theories. There may also be both developmental and interspecies differences in this regard. Young children and lower species may form expectancies by associative mechanisms, the former having yet to build any theories and the latter probably unable to build them; whereas adults and higher species may largely generate them from theory.

Experiments to decide among the alternate interpretations

Since the experiment that opened the way to the present theory cannot itself decide among the alternative interpretations, we will describe some other experiments that have more resolving power. No single experiment can be all things to all objections, but the proper combination of results from these experiments could decide the issue nicely. In one or two cases the results are already at hand, but unfortunately in many cases the tests are waiting in the rather long line of studies that remain to be done.

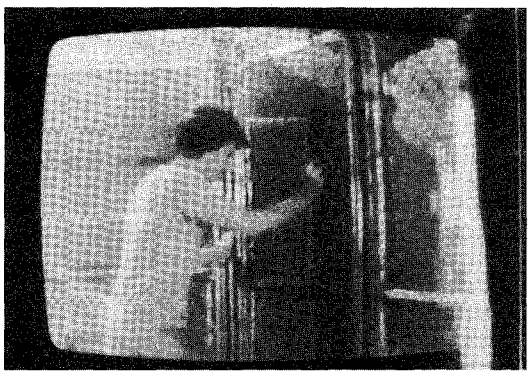
Problems not restricted to physical inaccessibility

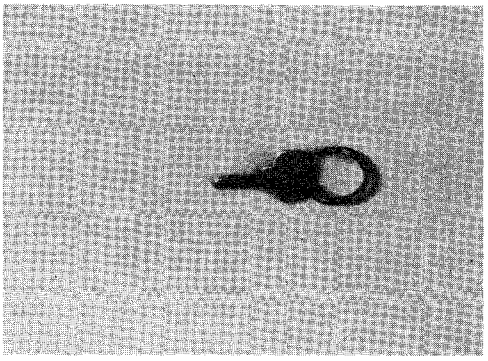
The first type of experiment broadens the definition of what constitutes a problem, increasing its scope, giving us much more room to work in. Notice that "problem" has so far been restricted to food (or some prepotent item) that is physically inaccessible. Compared to the human concept of problem, this is an almost unthinkable restriction. Problems, in the human case, can be instantiated in indeterminately many ways: by an intractable mathematical derivation, a car that will not start, a recalcitrant spouse, and so forth. Is the chimpanzee's concept of problem really as narrow as the existing experiments suggest? Or is the narrowness attributable to experimenters who have not tested the ape in ways that would disclose the abstractness of which he is capable?

NOTE

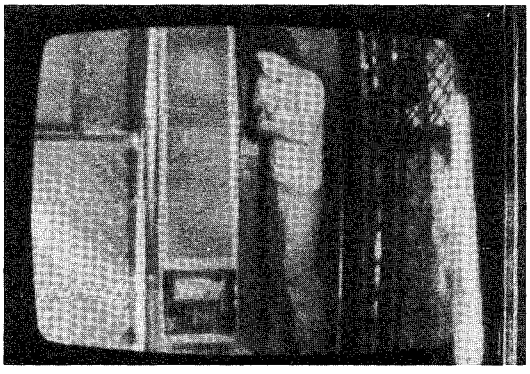
1. Empathy can be formulated in either of two ways. The chimp picks an alternative that describes what he would do (1) if he were in the actor's position, or (2) if he were a three-year-old child, a juvenile chimpanzee, a human adult, and so forth. It is not clear that the second formulation can be differentiated from theory of mind. To behave as a three-year-old child would presuppose a knowledge of what a three-year-old child knows, and this seems equivalent to making inferences about the knowledge of another.

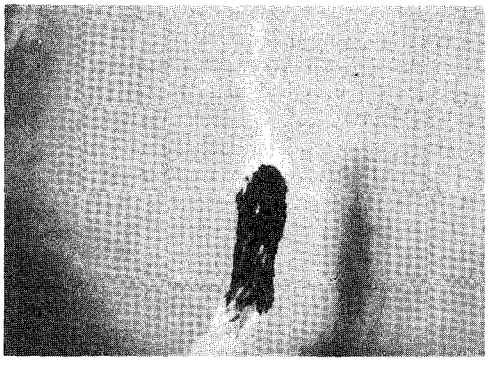
https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press



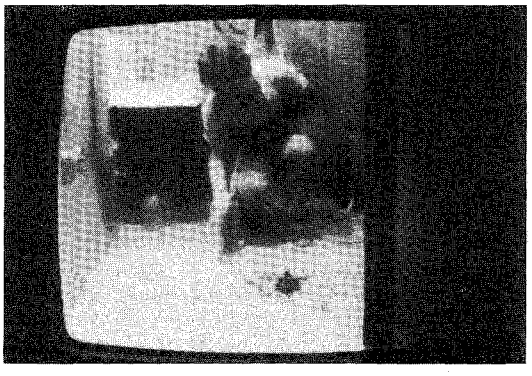


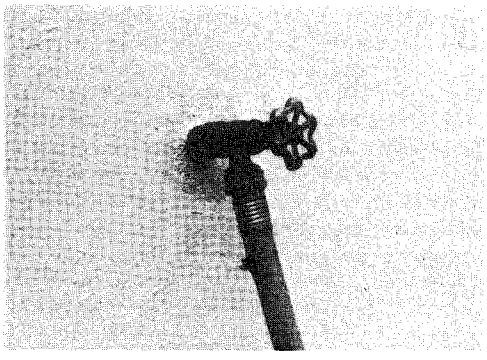
5



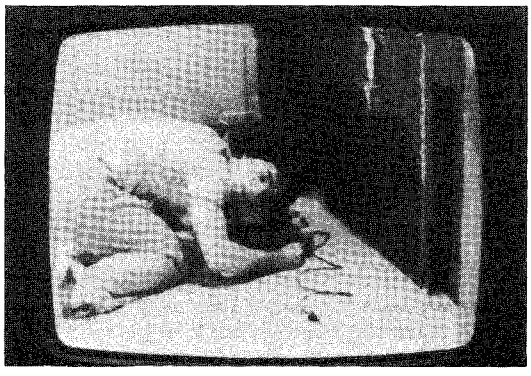


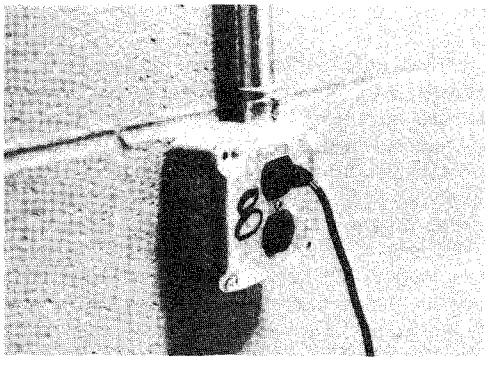
6





7





8

Figure 2. Photographic reproductions from the last 5 sec. of each of the four 30-sec videotaped problem scenes in Test 1 appear in the left column. In the right column are shown the still photographs of the correct solutions. For example, in Problem 5, the human actor struggles to escape from a locked cage, alternately grasping the bars of the cage and the padlock on the door, and Solution 5 shows a key. See text for fuller description.

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

To escape the usual definition of animal problem, we tested Sarah on four cases quite different from the usual variety. She was shown (1) a (human) actor struggling to escape from a locked cage, (2) a malfunctioning heater (as witnessed by an actor who glanced wryly at the heater, even kicked it a little, and at the same time shivered and clasped his arms to his chest), (3) an actor seeking to play an unplugged phonograph, and (4) an actor unable to wash down a dirty floor because the hose he held was not properly attached to the faucet. It is clear that in these cases "problem" has taken on a richer meaning. No longer a banana that is out of reach, "problem" now ranges from a human actor locked in a cage to a human actor "registering" cold, and thus suggesting that a heater has gone out.

Sarah was tested on these cases in the same way that she was tested on the first set. The last five seconds of the videotape depicting each problem was put on hold, and she was given pairs of photographs between which to choose (Figure 2). We ran this set of problems first with gross alternatives and then with substantially more discriminating ones.

On the first series the alternatives consisted of a key, an attached hose, an electric cord properly plugged into a socket, and a lit cone of paper (of a kind normally used as a wick to light the pilot). With alternatives of this kind, Sarah made no errors whatsoever. She paired the key with the locked-up human actor, the burning wick with the unlit heater, the plugged-in cord with the unplugged phonograph, and the attached hose with the disconnected one (Premack & Woodruff, 1978). Two of these cases – the nonfunctioning hose and the disconnected electric cord – could be accounted for by physical matching and are therefore not interesting as such. But two could not be accounted for in this uninteresting way: there is no physical match whatever between a key and an actor struggling to escape from a cage, nor between a flaming paper cone and an actor shivering and glancing wryly at a heater.

In the next series we presented the same problems, but now with greatly refined alternatives. Sarah was no longer required simply to choose among such grossly different alternatives as keys, hoses, cords, and flaming paper. Instead, she was presented with three versions of each of the four cases. For instance: key intact, twisted, or broken; hose (or electric cord) attached, not attached, or attached but cut; roll of paper unlit, lit, or burnt out. On this series she made one error in [twelve choices,][d] choosing the twisted rather than the intact key. This error was attributable to the quality of the photography. In the original 8″ × 10″ photographs, it was easy to see the difference between the intact and twisted key, but when we had the photographs reduced to 3″ × 4″ as an economy measure, it was difficult to detect the twisted key.

None of her answers in this second series could be attributed to simple physical matching. She could not simply match a hose to an unconnected hose but had to be able to discriminate an attached hose from one that was attached but cut off at the bottom; likewise for the electric cord. Similarly, she had to be able to appreciate the advantage of the burning paper over paper that was either not yet lit or had once been lit but had since burnt out. In brief, to pass these tests, she had to recognize not only the correct item, but the correct item in the correct state.

Could Sarah have solved these problems simply by choosing old or familiar scenes over new or unfamiliar ones? Even if we grant her visual memory of human quality, so that she could readily have distinguished old from new, she could not have used this distinction to differentiate between correct and incorrect alternatives. She could have rejected a few cases as unfamiliar, for example, the cut off hose and electric cord, but not the majority of cases. Hoses that are disconnected and electric cords that are not plugged in must have been familiar sights to her. Even more impressive, in the choice between the three conditions of the paper wick, the not-yet-lit wick and the burnt-out one must have been no less familiar than the lit one.

Notice that her success on this series was based entirely on observational learning. While she had often seen trainers play the phonograph, wash the floor, and the like, she had never carried out any of these acts herself. Moreover, the observational learning here is of a special kind. Each act, of course, is actually a complex sequence involving a series of acts, as in playing the phonograph: bringing the device into Sarah's room, plugging in the cord, turning on the machine, putting on a record, adjusting the head, and so forth. The acts were never segmented for her. For instance, we had never made a little ceremony out of plugging in the cord, one that would have pulled it out of the behavioral stream and didactically highlighted it. On the contrary, all these acts were part of the daily routine, so removed from being seen as pedagogic material that we never even called Sarah's attention to them, as one does routinely in teaching or training. Trainers came and went, lighting the heater, washing the floor, and the like. Sarah watched them from her cage, and in the course of watching, evidently segmented the acts herself.

This series established that the chimpanzee's concept of problem is not the restricted one suggested by earlier tests. And in all likelihood these tests do not begin to exhaust the possible abstractness of the animal's concept of problem. Yet even by themselves they show how nice the animal's physical knowledge can be. Of course we are not interested primarily in the animal's grasp of physical relations. But the fact that the chimpanzee has such a grasp makes possible a kind of test in which we are interested.

Modalities: "would," "should," and "would like"

When Sarah chooses a photographic alternative in experiments of the kind thus far described, she may intend it as an answer to any of three questions: What *would* the human actor do if in this situation? What *should* he do? What *would I like* to see him do? It is possible, of course, that she is not answering these, or any other questions, but is simply choosing the photograph she prefers. If we adopt that position, however, we must explain why her preferences bear the relations they do to particular videotapes. For, as a matter of fact, she does not choose a picture of a key, someone standing on a box, or anything else, on simply any occasion whatsoever, but exclusively on those occasions when the picture constitutes what we call the "solution" to the problem shown in the videotape. For this reason, it makes sense to try to understand Sarah's choices not as simply reflecting preference for pictures, but as answering one question or another.

It is clear that the human adult can operate in any of the above modalities, and that simply by asking him, "What would the actor do if in this situation?" "What should he do?" "What would you like to see him do?," we could obtain answers specific to the modalities. Not only could the human adult answer each of these questions, but he can apparently shift from one to the other with impressive ease. Changing from "would" to "should" to "would like" and back to "would" does not resemble shifting gears or moving pins in a plugboard. What seems notable, in part, about the human capacity for operating in different modalities is the evident ease with which the change is made (although admittedly this is based on impression, not actual study).

Can the chimpanzee distinguish among the modalities, operating first in one and then the other, shifting among them with greater or lesser ease? We cannot answer on the basis of the tests described so far, for they have been noncommittal with regard to these distinctions. It may be possible later to introduce markers that will make these distinctions explicit, but for the moment the tests have been no more discriminating than they need to be in order to yield data capable of deciding between competing theories.

[Moreover, for the time being, rather than attempting to introduce markers for "would/should" and so forth, we took a different approach. We returned to the problems illustrated in

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

Figure 1 and reproduced them with another human actor. The actor shown in Figure 1 is Keith, Sarah's favorite trainer. We also produced an identical series of four problems in which the actor was Bill (fictitious name), an acquaintance of Sarah's but one for whom she displayed no affection. In addition, we made a double set of alternatives for each problem. The "good" alternatives had already been made, and are those illustrated in Figure 1. We now made up "bad" alternatives for each problem, (the same ones for both actors). In one case, this alternative merely consisted of an inappropriate object: a rod that was too short to reach the bananas.][e] But in the other three cases, the "bad" alternative depicted a mishap that could have occurred to the actor in the course of attempting to solve the problem. The actor was shown stepping through the box, falling over the box, or lying on the floor with the cement blocks strewn over him. Of course, the point of the second series was to see whether Sarah's choice of alternatives would vary with the actor. Would she assign "good" alternatives to Keith, whom we had reason to believe she liked, and "bad" alternatives to Bill, for whom we had reason to doubt her affection?

We used the same procedure in these tests as in those described above, except that the trainer did not respond differentially to Sarah's choices but approved of all of them. In each of four sessions, Sarah was shown four videotapes, two of each actor in a counterbalanced order. A "good" and a "bad" alternative were presented with each videotape.

For the problems in which Keith was the actor Sarah chose the "good" alternatives eight out of eight times, but with Bill, only two out of eight [times.][f] If we designate a "good" photograph with Keith and a "bad" photograph for Bill as "appropriate," then the run of ten consecutive "appropriate" responses in thirteen trials that Sarah made on this test is [highly][g] significant ($p < .01$, Grant's test). On this test at least, Sarah's choices could be interpreted as answering the question, "What would I like to see happen to agents whom I do and do not like?" We cannot rule out, of course, that she regards Bill as less able than Keith, and that her choices answered the question, "What would happen to agents who were and were not competent?" But this seems even less likely than the first interpretation, which is already unlikely enough. Moreover, before we can seriously defend either interpretation, it is necessary to rule out a still simpler one. These data do not by themselves rule out that Sarah simply prefers the "good" photographs of Keith and the "bad" ones of Bill.

To determine whether or not Sarah's choices were affected by the intrinsic content of the photograph, independent of its relation to the videotape, we tested her again on the same material, but now with three alternatives for each videotape. These were the "good" alternative, the "bad alternative, and a second "bad alternative, irrelevant to the problem in the videotape. Each "bad" alternative was used equally often as an irrelevant alternative and was presented with the same pairs of relevant alternatives for each actor. Sarah was given sixteen trials, eight in each of two sessions.

Sarah chose the "good" alternative seven out of eight times for the problems in which Keith was the actor and only one out of eight times for those in which Bill was the [actor.][h] This is in keeping with her previous results. However, of the seven "bad" alternatives she assigned to Bill, three were irrelevant. This suggests that she is affected by the immediate content of the photographs, independent of the videotape with which they appear. This does not say, however, that she is not also affected by the relation between the contents of the photograph and the videotape. To assess the possible role of these two factors, we must be able to control or eliminate one of them. In the next step we managed to control her simple preference for the photographs per se.

The total frequency of her photograph choices in the previous two tests showed that she chose some more often than others, confirming the suggestion of preference at the level of the photograph. Her favorite "bad" alternatives were the prone actor strewn with cement blocks and the actor using the short rod (twelve choices), compared to the actor either stepping through the box or falling over it (three choices). We divided the "bad" alternatives into high and low preference on the basis of these frequencies and used this simple classification to control for preference.

This time, in retesting Sarah on the same material, we restricted the pairing of "bad" alternatives to those of comparable preference. As before, we gave her three alternatives with each videotape, a "good" one and two "bad" ones, but now each pair of relevant and irrelevant "bad" alternatives was of roughly equal preference. That is, we never paired a high and low preference "bad" alternative, but gave her only pairs in which both members were either high or low. In addition, we tested her only on the videotapes with Bill, for it was essentially only on those problems that she chose "bad" alternatives.

Even this relatively crude control was sufficient to reveal the effect of the relation between the content of the videotape and that of the photograph. She chose the "bad" alternative ten out of twelve times, nine of them relevant and only one irrelevant ($p < .05$). Incidentally, the role played by preference in these results is hardly unique. Preference can affect not only presumptive answers to implicit questions that are contained in visual material but also explicit answers given to linguistic questions. Preference has been shown to affect both the comprehension and production of language (Premack, 1976). In general, the role of nonverbal factors in both verbal and nonverbal tests deserves wider recognition.

Before closing this section, a word in Sarah's "defense" may be advisable. Because she favors pictures of untoward outcomes for an actor whom we suspect she dislikes, we cannot be certain that she would choose to inflict the same actual outcomes on the person in question. She might, but we do not know; and it is not an easy question to answer. Finally, regardless of how these results are ultimately interpreted with regard to the modality question, notice that they provide no support for the empathy view. Sarah is not simply choosing alternatives in keeping with what she should do if she were in the actor's position. Her choice of alternatives depends on who the actor is.

Agent-specific knowledge and belief

Although in the tests above, Sarah's choice was affected by the actor's identity, we do not know whether she could use the actor's identity in more demanding ways. Could she assign different purposes to different agents? Could she assign different knowledge or beliefs? The former is easier to picture than the latter, for motivational states seem more primitive than cognitive ones (though, to be sure, separating the two kinds of states is problematic).

Suppose Sarah were shown both an adult and a young child facing a relatively complex device such as an unlit heater. She could then be called upon to indicate by her choice of photographs which of the two agents could light the heater, how they would proceed, and whether one of them would be more likely to make errors (such as applying a lit match to the wrong place or an unlit match to the right place, applying a lit match for too short a time, etc.). If Sarah imputes not only purpose, but also knowledge (and her inferences about knowledge are even approximately correct), then the alternatives that she chooses for the adult and for the young child should differ appreciably.

Of course, we cannot expect Sarah to make accurate inferences about knowledge and belief without appropriate experience. She must know something about the complexity of heaters relative to the competence of adults and children. In one sense, her success on problems of this kind might be of limited interest. If she observed adults and children succeed and fail respectively on a particular task and then chose photographs in keeping with these

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

observations, we would be relatively unimpressed, having been jaded by some of the more mystifying results we had already seen. Performance of this kind would be of interest mainly for demonstrating the animal's ability to recognize correspondences between "real" experience and two-dimensional representations of that experience. This ability is surely not universally distributed over species. But in the present tests, rather than studying that capacity, we are taking it for granted and using it to address other questions.

[Results will be of interest only if there is an appreciable difference between Sarah's experience inside and outside the test situation. In the ideal case, we consider it possible to explain her success in a novel situation by claiming her choices were derived from a theory, in the same sense that one claims sentences are derived from a theory of grammar. In this case, however, instead of a grammar, an individual would have a system of imputations about: the knowledge and beliefs, the inferences, the problem-solving ability, and so forth, of another individual. Such a theory is not likely to be elegant or even accurate when found in the child or chimpanzee. Certainly the theory will reflect systematic errors or biases peculiar to the operating characteristics of the species. We assume that the theory would be learned, in the same sense that natural language theory is learned; but what precisely is that sense?

Are theories built up out of such items as "the bait is on the right side," "triangle is correct," "light and food are positively correlated?" These rather typical examples of laboratory learning are not helpful, either in clarifying the kind of information that goes into making a theory, or in showing how such information would be learned. For instance, in stating relations between classes, theories presuppose a knowledge of class properties, and ability to identify class members. But laboratory exemplars of learning do not illustrate how classes are learned, or help us to understand why the individual forms particular classes or draws particular generalizations, when in any situation he can learn indeterminately many classes or generalizations (cf. Premack, 1973).]

To determine whether or not Sarah would choose different photographs for different agents confronting the same problems, we made two series of videotapes, one with a four-year-old female child and another with a female adult (neither of whom were known to Sarah) as the actresses. Each series consisted of six tasks: object match-to-sample, shape match-to-sample, a memory task, and completion, causal inference, and spatial matching tasks. We photographed a correct and an incorrect solution for each videotape with both actresses. For example, on the memory problem the videotape showed one of the actresses watching the trainer place an apple under one of three containers; the correct photograph showed the same actress pointing to the correct container while the incorrect photograph showed her pointing to the incorrect container. All photographs were alike in the important sense that they showed not only the correct outcome (e.g., a ball placed with another identical ball, in the case of object matching) but also the appropriate agent carrying out the act.

Sarah was tested as before, one trial on each problem, with the order of the problems for the two agents counterbalanced over trials. On each videotape she was given two photographs, one depicting a correct act and another an incorrect one, by the appropriate agent. The trainer approved of all of her choices.

Sarah chose the correct photograph on ten out of twelve trials, five for the adult and five for the child. Unfortunately, this negative outcome is rather indeterminate. First, Sarah has had little experience with children (and none in recent years), whereas she has had appreciably more with young chimpanzees; new tests will contrast young and old chimpanzees as actors, as well as people versus chimpanzees. Second, and probably more important, the six tasks are not well graded in terms of difficulty. (They are listed above in the order of the presumptive difficulty; even though Sarah now has a bit more trouble with some than with others, in future studies it will be possible to include more sharply differentiated tasks, such as reassembling diagonally bisected photographs, which, unlike horizontally or vertically bisected ones, she is still unable to reassemble.) Indeed, four-and-a-half-year-old children can (as we recalled in retrospect) do all six of them. Thus, if we ignore Sarah's two errors, we could regard her choice of alternatives as an accurate assessment of the competence of both the adult and child. We propose this facetiously, of course, and will correct for the indeterminacy in future tests.

Embedded videotape: observing a viewer observe. An especially discriminating test, which we have not yet used, is based on the embedded videotape. The virtue of this test is that it removes all possible unclarity as to what is the basic question that we are trying to answer. Sometimes it may seem that our question is merely "what does the chimpanzee (or child) know about the world?" Information of this kind can be of interest, of course, but only as a preliminary to a deeper question. The embedded videotape makes clear that what we are testing is not "what does the subject know about the world," but rather "what does the subject know about what someone else knows (guesses, believes, thinks, etc.) about the world?"

To make this kind of test we prepare a videotape in which an observer O is shown watching a participant P. We use two cameras in making this videotape, for P is not acting in O's immediate world, but is an actor in a videotape. Thus O is essentially playing Sarah's role, that is, watching an actor on television; and Sarah, in using this tape, is watching O watch a videotape of P.

On the embedded videotape, P is shown locked in a cage, desperately trying to select the right key from a large assortment as a lion slowly edges forward. As O watches this scene, the videotape of P is put on hold, and O is presented with two photographs. In one, P finds the right key and escapes in the nick of time; in the other, P is not so lucky and succumbs to the lion.

Sarah's videotape is put on hold at this moment, and she is given two photographs in the usual way, one showing O pick the photograph in which P escapes, a second in which O picks the photograph in which P is not so fortunate. Sarah's task is the usual one, to choose between the alternatives.

Sarah is given this test not only with the actors O and P, but for many pairs, for each of which she has the same information: does O like or dislike P? A wide variety of contents could be used to provide this information to Sarah. To assure that we affect what the animal knows about each agent's attitude toward the other, rather than her attitude toward them, we do well to avoid actual trainers and use instead anonymous actors in videotapes. In this way we can also avoid content that is at all like that of the test itself; then the animal could not be predicting simply that O would do something to P that was part of the original evidence on which it based its inferences about O's attitude toward P. (Moreover, recall that in the test the animal does not predict what O will do to P, but only what outcome O would choose for P when shown a videotape of P.)

If, for those cases in which Sarah has been shown that O1 dislikes P1, she selects photographs showing O1 assigning untoward outcomes to P1, and for the opposite case, photographs showing O2 assigning benign outcomes to P2, we can conclude that she believes that O1 dislikes P1, and that O2 likes P2. In other words, we can assume that Sarah imputes states of mind to the agents in question, states consisting of either positive or negative atitudes.

The embedded videotape is, of course, not restricted to the present material but can be applied to any sort of content. For example, P, who can be either a child or an adult, is shown being given a standard conservation test. O, who can also be either a child or an adult, is showing observing the videotape that shows the testing of P. O is asked to choose between photographs that show P making a correct conservation judgment in one case and

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

an incorrect judgment in the other. Sarah's task is to choose the alternative that O would choose. In doing so, Sarah must decide not merely, "are children or adults capable of conservation," but rather, "do children or adults know that children or adults are capable of conservation?" That is, correct performance in this case would show not merely knowledge of what an adult or child can do, but knowledge of what a child or adult knows a child or adult can do.

The chimpanzee as learning theorist. Any animal who ascribes different states of knowledge to naive and experienced organisms must know about either learning or maturation: some process that will account for the transition from ignorance to knowledge. So far, we have asked "What mental states does the animal infer?" But with appropriate experiments we can also ask "What process does the animal infer as accounting for the transitions from one mental state to another?" Should the animal prove to assign different knowledge to the child and the adult, what kind of experience or process might the animal believe changes the child (who cannot properly reach inaccessible items or restore nonoperative devices) into the adult (who can readily do all these things)? What must the chimpanzee observe of a child and an adult before he assigns different general competences to them? Besides these questions, we may ask the phylogenetic question, "If the animal regards the adult as a transformed child, could he also be seeing the human as a transformed chimpanzee?" Perhaps in comparing species, he will reject all possibilities of transformation (thus contrasting with our own assumptions about evolutionary transformation). All these questions remain to be answered. They would follow quite directly should it turn out that the chimpanzee can recognize not only different mental states, but also the role of developmental and organismic variables in accounting for differences among mental states, and in some cases for the transition from one mental state to another.

Epistemic states: a program for their experimental investigation

The distinctions between *know* versus *guess* on the one hand, and *truthfulness* versus *deceit* on the other, can have great practical weight in human social affairs. Your choice as to whether or not to accept information from a potential informant is often filtered through these distinctions. For instance, you are lost in a city and seek directions. But something in the manner of your informant suggests that he is merely guessing, or that he is only pretending to know when he really does not, or that he knows very well, but is misleading you. You may also encounter the informant who relieves you of the burden of assessing his mental state, for he knows that he is guessing and tells you as much. In all these cases, you are likely to move on, even at some risk, and to look for the acceptable informant. The ideal informant is one whom essentially you could substitute for yourself without loss: he knows as much about the matter in question as you would know if you had had the benefit of his experience, and his account is impeccable. These distinctions play a vital role in human communication; we accept or reject information, depending on our assessment of them. Unfortunately, whether they play any role in the communication of other species is still largely unknown.

As yet we do not know whether Sarah distinguishes between *guess* and *know*, or between *pretend* and *real*. It may be of value, however, to describe how we are finding this out, if only to show how amenable mental states are to study.

To determine whether the chimpanzee distinguishes between *guess* and *know*, we can use either of two approaches: sorting or language markers. In both cases we select the clearest possible exemplars of the two states and then use them either to introduce the markers *guess* and *know*, or simply to establish two categories (differentiated by their location in space). Evidence for the animal's grasp of the distinction would lie in the transfer performance data, in terms of either sorting new cases correctly, or applying proper markers to them. The sorting and language approach are not basically different. One would choose the latter only if the animal had already been taught other aspects of language, for then he might be able to combine the new words *guess* and *know* with existing ones, forming strings that he could not otherwise form.

"Guess" versus "know"

In an experimental program currently in progress we use three different contrasts to exemplify *guess* and *know:* a naive and a sophisticated rat at a choice point; a human actor drawing marbles from an opaque and a transparent urn; a judge in a recognition experiment identifying items on the basis of viewing times that are hopelessly brief in some cases, comfortably long in others.

In the first case, the animal is shown a naive rat at a choice point – its vacillation, long latency, and frequent errors. Later in the videotape, the same rat is shown at a more advanced state of learning, choosing correctly and doing so with a short latency. The early and later behaviors of the rat are intended to exemplify *guess* and *know*, respectively.

In the second videotape, the animal observes a human actor drawing black and white marbles from an urn. Sarah is shown that the urn contains an equal number of black and white marbles, that the urn is opaque, and that the actor cannot see what he is drawing. Later in the tape she is shown a transparent urn, and that now the actor can readily see the marble he selects. In both cases, the actor predicts which color marble he will draw, making his predictions by pointing to either a white or black card before choosing a marble. Both the actor's choices and predictions in the two cases are taken to exemplify *guess* and *know*, respectively.

In the last case, Sarah is shown a human actor looking at the faces of different people through a porthole, trying to identify them. On some trials, the judge is given only a glimpse of the person, on other trials, a long inspection. The videotape gives Sarah a side view of the proceedings, enabling her to see both the judge sitting at the desk and the target person behind the screen. As a result, Sarah always knows who the target is. She is also shown the view that the judge has of the target individual, and can see that this view is sometimes painfully brief, at other times comfortable long. On each trial, the judge chooses one of the three photographs placed before him. When the judge is correct, a light below the porthole goes on; otherwise it remains dark. Sarah is in a good position to learn the significance of the light: she can observe the correlation between the individual who stood behind the screen and the photograph the judge chose.

If Sarah is successful on either the marker or sorting task, can we be certain that she is truly distinguishing between *guess* and *know*, as opposed to merely distinguishing between accurate and inaccurate prediction or choice? The recognition experiment can make a nice contribution to this problem. We can require Sarah to evaluate only those cases in which the judge was successful. Those cases are identified by the agreement between the individual Sarah saw behind the screen and the photograph the judge chose, plus the onset of the feedback light. If Sarah can differentiate between those cases in which the viewing time is impossibly brief versus those in which it is comfortably long, then she cannot be discriminating merely between successful and unsuccessful prediction, for in all cases the prediction is successful. Neither can she be discriminating short versus long viewing time, for these parameters have nothing in common with either transparent and opaque urns, or with vacillating and unhesitant rats. *Know* and *guess* are instantiated by indeterminately many, physically diverse conditions, and the question is, Can Sarah discriminate among them?

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

"Pretend" versus "real"

Pretend may be yet another state of mind that the chimpanzee will ascribe to himself and others, as many anecdotal claims have suggested. We now have laboratory evidence that points in this direction. Although the evidence is not definitive – more than one test will be needed – it is a step beyond the anecdote. The test is quite simple. A room is divided by a mesh partition. On one side is a pair of containers, one of them baited, and on the other side, cut off from the containers by the mesh partition, a chimpanzee. The chimpanzee knows which container is baited, having observed the baiting, whereas the trainer who stands on the side of the containers does not know. The animal can obtain the bait, however, for if the trainer is able to judge from the animal's behavior which container is baited, and can thus choose correctly, the chimpanzee is given the bait.

In a second version of this experiment, the trainer is villainous or selfish, for when he succeeds in choosing the correct container, he keeps the bait for himself. The two types of trainers wear highly discriminable costumes, so that the animals can know which condition obtains. We have combined a reversal of roles with these two basic conditions, so that the chimpanzee has served as a "receiver" on some occasions and as a "sender" on others. As receiver, the animal was uninformed (as was the trainer when he played this role), standing by the containers, waiting for the informed sender on the other side to direct him to the baited container.

True to form, the villainous trainer was selfish not only as receiver but as a sender as well, for as sender he invariably pointed the animal to the incorrect container. The benevolent trainer was equally consistent; as receiver, he not only gave the animal food, but as sender, he consistently pointed the animal to the correct container.

A simple strategy for coping with this situation is to tell the truth to the benevolent trainer and to take his guidance as truthful; conversely, be untruthful to the selfish trainer and take his guidance as untruthful. Lying or untruthfulness can take two forms (since the two containers constitute a mutually exclusive dichotomy): false negatives or false positives. In the former, one simply withholds information, that is, one does not make any of those responses that enable a receiver to judge which container is baited. In the latter, one directs the receiver to the incorrect container, either by using the very same response one uses to tell the truth, or by a response different from the one used for truth telling (which, of course, would be a detectable, and hence inadvisable strategy).

With four sexually immature chimpanzees who have been tested and retested on this problem in a longitudinal study of nearly two years' duration, we have at this time all possible outcomes (Woodruff & Premack, in preparation). The oldest animal is a successful liar in both production and comprehension; the youngest (but also probably the smartest) lies in comprehension but not yet in production; the one male lies in production but not in comprehension; and the fourth animal, who is a kind of social isolate, does not lie in either modality. (By lying, we mean false positives: all animals have shown false negatives to some degree and this form of behavior typically precedes the false positive.)

Although we have spoken of "lying," we do so for convenience. The observed behavior may well be a precursor to lying, but we know too little about the inferences the animals may be making to be assured that they are indeed lying. What we would need to know would be that the animal believes that the selfish trainer knows which containers is baited, and is purposefully directing him to the unbaited one. If we could be sure of that, we would be justified in calling the animal's false positives lying. We are testing that now in the following way.

The animal is allowed to make a choice from among 100 opaque containers placed before him. Before he chooses, he is told which one to choose, in one case by a liar (L) and in another by a well-intentioned fool (F). The consequence of this advice for the animal is about the same. L, who knows which one of the 100 containers is baited, never directs the animal to the correct one; F does little better. There are 100 containers, and since the F is guessing, he is seldom correct.

Despite the fact that F serves the animal no better than L, there is the possibility that the animal will nonetheless react differently to them. First of all, there is the opportunity to observe that the two parties do not have the same knowledge about the containers. The baiting is carried out in a room adjacent to the animal's. And through a window, the animal can recognize the 100-container device that is subsequently brought to his room. He can see that baiting is going on, but not which container is baited. He can also see that trainer L, identified by a prominent red hat, is invariably present during the baiting.

In contrast, trainer F, identified by a green hat, is never present during the baiting. Indeed, while the baiting is going on, F is standing at the rear of the animal's room, with even less chance than the animal to know which container is baited.

Once the device is located in the animal's room, it is of course exactly L and F who offer the animal advice: L on some trials, F on others. What we are interested in is the animal's response to L and F. Will he dislike L while tolerating F? He may dislike them equally for, as we have seen, their advice is about equally bad.

However, he might regard them differently, for, from the human point of view, only L is a liar, that is, only L knows which container is baited and is intentionally misleading him. F is simply uninformed. As the animal can see, F is never present during the baiting, and cannot therefore be accused of deliberately misleading him.

Why should F participate at all? We might make the whole situation more plausible by putting it in a social frame. Normally the animal is guided by a benevolent trainer, who not only baits the 100 containers and carries them into the animal's room, but never fails to direct the animal correctly. That is what usually happens. But on some trials the phone rings and he is called away. Then L and F, who are almost always at hand, essentially fill in for the benevolent trainer, L with lies and F with well-intentioned misinformation.

We are running this experiment now, with the hope that the animal may make a distinction, comparable to our own, between someone who is guilty by intent and someone who is guilty by ignorance. We had been encouraged by an earlier finding. Although the young animals have shown little open animosity toward the selfish trainer, this was not so with Sarah. After only two or three experiences with the selfish trainer, her aggressive displays toward him were such that we thought it dangerous to continue the experiment. Toys and other objects lying about the cage, many of them sharp and hard, sailed out under the mesh at dangerously high speeds, narrowly missing the selfish trainer. But is this Sarah's response to a liar, or simply to anyone who misguides her? Moreover, in the first experiment one might question whether Sarah knew that the selfish trainer was a liar. Since she never actually saw him watch the baiting, which was done out of her view, how could she know that he knew which one was baited? On the other hand, there were only two containers, and no one can *consistently* guide you to the empty one without knowing which one is baited (but does Sarah know that?). In addition, Sarah's experience with trainers was like the child's with teachers: teachers and trainers are always informed, they always knew the answer (cf. Milgram, 1974). So if a trainer has consistently directed you to the wrong container, he must be lying. In any event, the present experiment will distinguish between the two alternatives: Sarah responds with hostility toward anyone who misguides her; and she responds with hostility to a liar: someone who misguides her intentionally.

Lying, however, is only one form of *pretending*. Like *guessing* or *knowing*, pretending can take many forms. The interest in this case is accordingly also like that in the others: can Sarah label or group together the several forms of pretending despite differences in their form? To generate some further cases of pre-

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

tending, we can return to the Köhler type problem. Even in this simple case, "problem" is instantiated by two factors, a physical and a psychological one: food is out of reach (physical factor) and an actor is trying to get it (psychological factor). Combining these factors of course yields four cases. In one, food is out of reach and the actor is trying to get it; positive on both factors: a full-fledged problem. In a second, food is neither out of reach nor does the actor try to get it; negative on both factors: not even the possibility of a problem.

However, pretend-problems can be instantiated by all combinations of the two factors, because the value of the psychological factor need not be real, but can be simulated. For instance, an apple hangs just above the actor's head, and yet somehow he "cannot" reach it. Or a hungry actor shows no interest in the inaccessible fruit, until a second party leaves, whereupon he directs his full attention to it. Or a nonhungry actor jumps repeatedly for an inaccessible apple because a moment earlier he had noticed Burt, a chimpanzee escape-artist, sneaking along the wall heading for the door; the actor's sudden intense interest in the apple fools Burt, who runs by and is easily grabbed.

But Sarah may be less easily fooled, especially when shown a videotape of the whole proceedings, enabling her to see both the actor's and Burt's points of view. When Sarah is tested on these cases, with either a sorting or marker approach, her success on the transfer material would demonstrate her ability to distinguish between genuine and simulated problem solving. Of additional interest would be the nature of the information Sarah must be given in order to draw this distinction correctly. For instance, if she is shown *why* an actor is simulating, will she be helped, as we would be, to "see through" the behavior and detect the simulation?

Self-knowledge. This same approach can be used to determine not only whether the chimpanzee attributes mental states to another, but also whether she attributes them to herself. Does Sarah know when she is truth-telling or lying, when her decision is based on knowledge or when she is merely guessing? Just as she can be required to apply markers ("truth/lie," "know/guess") to the acts of others, can she be required to apply them to her own acts? Will she be as successful in learning to apply markers to herself as she is in learning to apply them to others? Comparisons of this general kind can be used to answer the perennial question of how self-knowledge grows. Does it develop primarily from observation of others, from the observation of oneself, from introspection, or from subtle interactions, among all these cases, and differently at different developmental stages?

Two parties that would hold a noninferential view: positivists and young children

We have urged that when the adult chimpanzee watches the film showing the trainer struggling to reach inaccessible bananas, he makes sense of the scene by imputing a purpose to the trainer. The reader, we suspect, will find it so natural to read the scenes in this same way that he may ask, "who would not read them in this way?" We can think of two groups, one sophisticated and the other naive. Positivists and other highly trained parties would not read them in this manner. Having learned to distinguish data from inference, they will, we think first read the scene in an inferential manner, like everyone else, but then go on to inhibit or suppress the inference. Having inhibited their natural tendency, they could then provide a description roughly along these lines: "Keith is in a cage-like area. Bananas are above his head. Keith is jumping up and down. He extends his arm above his head, in the direction of the bananas." The exact form of the description is unimportant; it is simply the lack of inference to which we wish to call attention.

The second group who might give a "description" quite surprisingly like the positivist's would be young children, so young that they did not understand the scene. Since the material in question is very simple, we may find it necessary to encourage the children's failure to understand by chopping up the scene and presenting it out of sequence. This makes understanding difficult, and in some cases even impossible. In studying children's comprehension of picture stories, we have found that older children, and of course adults, are able to make sense of picture stories even when the pictures are out of order, but that children, younger than about four cannot (Poulsen, Kintsch, Kintsch and Premack, in press). Young children's descriptions of out-of-sequence pictures are very reminiscent of the kind of description we have ascribed to the positivists. The young child's account does not attribute any purpose or intention to the actors. Indeed, it is exactly the absence of these inferences that makes the description senseless and disjointed; for example, "a man, a dog, a tree, the dog is running." In normal perception, the elements are joined together by imputing mental states to the actors; for example, "the dog is afraid of the man and is running away," and so forth. It is inferred mental states that "organize" scenes, holding together the disparate elements into which scenes otherwise threaten to dissolve.

The important point here is that assigning mental states to another individual is not a sophisticated or advanced act, but a primitive one. Only on two occasions are the inferences not found: when there is not enough understanding of the scene to permit the inference, as in the young, confused child, or when the inference indeed occurs, but is quite deliberately suppressed, as by a sophisticated adult who, having been taught the differences between data and inference, elects on this occasion to give what he calls an objective "description."

We think that an analogy can be drawn with causal inference, in which the conditions seem much the same. Whenever A precedes B, the belief that A causes B is, we think, the primitive one. This is a belief that would occur under all conditions except the two described above. The sophisticated human adult has been taught to demand more than co-occurrence as a basis for assuming causality. He will block the primitive inference and impose further tests. Are there any occasions when A occurs and B does not? When A does not occur and yet B does? When B occurs and A does not? When B does not occur and yet A does? If the answers to these four questions are appropriate, the sophisticated adult will drop the barrier and allow the primitive inference of causality to go through. The actual content of a sophisticated causal inference may not differ importantly, however, from the content of a primitive one: what differs is the process whereby the inference is reached.

Concluding remarks

In assuming that other individuals *want, think, believe,* and the like, one infers states that are not directly observable and one uses these states anticipatorily, to predict the behavior of others as well as one's own. These inferences, which amount to a theory of mind, are, to our knowledge, universal in human adults. Although it is reasonable to assume that their occurrence depends on some form of experience, that form is not immediately apparent. Evidently it is not that of an explicit pedagogy. Inferences about another individual are not taught, as are reading or arithmetic; their acquisition is more reminiscent of that of walking or speech. Indeed, the only direct impact of pedagogy on these inferences would appear to be *suppressive,* for it is only the specially trained adult who can give an account of human behavior that does *not* impute states of mind to the participants. All this is to say that theory building of this kind is natural in man.

Are we to believe, however, that we are the only species in which it is natural? Our series of comparative studies is devoted to this and related questions. Although here we have talked only

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press

about the chimpanzee, the same videotapes, with only a few exceptions, are used with two other populations: normal and retarded children. The group of studies is designed to answer the questions: Does a chimpanzee make inferences about another individual, in any degree or kind? Are at least some retarded children deficient in specifically this form of theory building? What is the developmental course of such theory building in the normal child?

Since there are many points to cover, the data are only now appearing, and we cannot yet describe the distribution of mentalistic inference, cross-specifically or developmentally. It is all the more difficult, therefore, not to venture just a little further speculation: Of all possible guesses, we find the most compelling one to be that inferences about motivation will precede those about knowledge, both across species and across developmental stages. Not even the chimpanzee will fail tests that require him to impute *wants*, *purposes*, or *affective attitudes* to another individual, but he may fail when required to impute states of knowledge. For instance, the ape may have no difficulty deciding that in a certain situation the child will seek its mother's arms, whereas the adult will seek food. But he may have appreciably more difficulty in deciding that the adult will seek food by masterful acts: appropriately opening a locked chest, stationing a ladder below a perch, and so forth; whereas the child will seek its mother by decidedly less masterful means.

The predicted results will leave us with a tricky question. Does the chimpanzee simply fail to impute knowledge to others, all his inferences being of the motivational variety, or does he indeed impute such states, but poorly, making gross errors in the *content* of the knowledge that he imputes? We may make some progress with this question once we know definitively whether or not the chimpanzee distinguishes *guess* from *know*. For if he makes this distinction, while at the same time distinguishing poorly between the acts children and adults would bring to bear on a common problem, this cannot be simply because he does not impute states of knowledge. It would rather seem to be a deficiency in the accuracy with which he discriminates between different possible contents of knowledge.

On the other hand, the ape may be incapable of differentiating between *guess*, and *know*, *doubt* and *believe*, and so forth. Then it may make more sense to consider that he does not make inferences outside the motivational realm. The immature child may resemble the ape in this regard. We may even wish to consider that the distinction between motivation and cognition, between *want* and *know*, is deeper, more biologically real, than that implied by textbook headings.

Having decided that behaviorism is unnatural because it requires suppressing primitive inferences, whereas theories of mind are natural, can we conclude that mentalism is, therefore, preferable, and more likely to lead to valid theories? Regretably, there seems to be no way of answering this question that will grant major catharsis to either camp. On the one hand, beliefs cannot be extolled simply because they are natural; naturalism does not guarantee validity. Indeed, some quite interesting recent work is concerned exactly with deciphering the miscalculations to which human reasoning is naturally prone (Tversky & Kahneman, 1977). Certainly, holding causal inferences in check, until all four cells in the contingency table (described above) have been tallied, is a highly unnatural but excellent idea.

On the other hand, if mental theories are indeed natural, this fact must have untoward consequences for behaviorism. After being shown that not only man but also apes had theories of mind, suppose a behaviorist were to reply, "Yes . . . and they are both wrong." Would this save behaviorism? We think not, for to admit that animals are mentalists compels the admission that behaviorist accounts of animals are at best profoundly incomplete. Moreover – and we add this with more than facetious intent – it would waste the behaviorist's time to recommend parsimony to the ape. The ape could only be a mentalist. Unless we are badly mistaken, he is not intelligent enough to be a behaviorist.

ACKNOWLEDGMENT

The research reported here was supported by National Science Foundation grants BMS 75-19748 and BMS 77-16853, National Institutes of Health grant 1-PO1-HD-10965, and a facilities grant from the Grant Foundation. We thank Sarah's trainer, Keith Kennel, for his collection of the data, and A. J. Premack for helpful comments on an earlier version of the manuscript.

REFERENCES

Köhler, W. *The Mentality of Apes.* London: Routledge and Kegan Paul, 1925.

Milgram, S. *Obedience to Authority: An Experimental View.* New York: Harper & Row, 1974.

Poulson, K., Kintsch, E., Kintsch, W., and Premack, D. Children's comprehension and memory for stories. *Journal of Experimental Child Psychology,* in press.

Premack, D. Cognitive principles? In: F. J. McGuigan and D. B. Lumsden (eds.), *Contemporary Approaches to Conditioning and Learning,* pp. 287–319, 1973.

Premack, D. *Intelligence in Ape and Man.* Hillsdale, N. J.: Lawrence Erlbaum Associates, 1976.

Premack, D., and Woodruff, G. Problem-solving in chimpanzee: Test for comprehension. *Science.* 202:532–5. 1975.

Tversky, A., and Kahneman, D. Causal schemata in judgments under uncertainty. In: M. Fishbein (ed.), *Progress in Social Psychology.* Hillsdale, N. J.: Lawrence Erlbaum Associates, 1977.

Woodruff, G., and Premack, D. Intentional communication in chimpanzee: the development of deception. In preparation.

EDITOR'S NOTE

The Commentary process makes it impossible to alter a target article once it has been circulated to commentators (as indicated in the Instructions to Authors and Commentators). The following errata are accordingly listed separately. The commentators saw only the original text as it appears above (square brackets and superscripts were added in proof):

[a]reaching out with a rod

[b]six

[c][delete]

[d]eight choices ($p < .05$),

[e][types of problems illustrated in Figure 1 and made eight new videotapes, using new props and one of two human actors in each scene. One actor was Keith (the person shown in Figure 1), Sarah's favorite trainer, and the other] was Bill (fictitious name), an acquaintance of Sarah's, but one for whom she displayed no affection. We also made 16 new photographs, one "good" alternative (similar in theme to that shown in Figure 1 for each problem) and one "bad" alternative for each actor in each problem. In one case, this "bad" alternative merely consisted of an inappropriate object: a rod that was too short to reach the food.

[f]times ($z = 2.58$, $p < .01$).

[g]also

[h]actor ($z = 2.50$, $p < .05$).

[i][delete]

https://doi.org/10.1017/S0140525X00076512 Published online by Cambridge University Press