This article originally appeared in:

Dennett, Daniel C. "Beliefs about Beliefs" (commentary on Premack, et al.). Behavioral and Brain Sciences 1 (1978): 568-70.

Beliefs About Beliefs [P&W, SR&B].

by Daniel C. Dennett*
Department of Philosophy, Tufts University, Medford, Mass. 02155

Because of its intrinsic interest - indeed its fascination - it is easy to lose track of the point of this kind of research. Getting children to talk takes on the aspect of sending a man to the moon. Suppose you succeeded. Then what? Presumably behaviorists would have to claim to be unimpressed, as unimpressed as they are by the verbal abilities of - themselves, for instance. So suppose we grant for the sake of a superannuated argument that

in principle a suitably complex version of behaviorism can "handle" all ape behavior (and all human behavior too). That version of behaviorism will of course be scarily distinguishable from more mechanic materialism - with mico- events in the brain being viewed as responses, for instance - and there is scant reason to oppose that creed, at least at this stage of our knowledge. The issue that remains is, on a first pass, how fancy a cognitive structure is required in practice to predict a chimpanzee's behavior. That is, granting that in practice it is desirable to intentionalize our account of chimpanzees (by attributing beliefs and desires, or belle like states or desire like states (Dennett, 1971, 1976), which beliefs and desires will it be useful, predictive, illuminating to attribute? In the present instance, will we find it valuable to attribute second-order beliefs and desires - beliefs and desires about the beliefs and desires of others? If so, then chimpanzees have a theory of mind in the requisite sense, for they use the concepts of belief and desire (or concepts importantly analogous) in their own action governance. If they turn out to have human like theories of mind, they will have use of even higher order intentional attributions: they perhaps believe someone wants them to believe something, or want someone to believe they want something, and so forth. But how can these suppositions be put to the test?

I think the issue is analogous to the current controversy about mental images. What the growing literature on mental images shows is that whatever it is to which we may in the end "reduce" mental-image talk, there can be no doubt that there is a level of description of the phenomena at which imagistic characterizations are perspicuous because they are richly predictive of a surprisingly wide variety of behavioral effects. Talking of mental images may be a *façon de parler*, but it is no "mere" *façon de parler*, because talking the talk (quite) literally keeps on leading to confirmed predictions. This is undeniable even if it is also true that talking about mental images is itself in dire need of explanation - and even of ultimate 3-D pictures in the brain.

What must be shown by Pramack & Woodruff, analogously, is that imputing a theory of mind to chimpanzees (whatever that comes to literally, in the end) is richly predictive. As P&W note, any single test, however consonant its results with the theory-of-mind hypothesis, can be given a deflationary redescription by associations et al. What one wants is a panopoly of results elegantly predicted by the theory-of-mind hypothesis and only predictable with the aid of ad hoc provisions by its competitors. P&W do not yet have these results, as they grant, but while the experiments they are now undertaking would favor their hypothesis if the results were positive, they seem somehow slightly off target. P&W are searching for evidence that chimpanzees have expectations of the behavior of others that are better explained by supposing that they are (tantamount to) predictions derived from the chimpanzee's beliefs about the beliefs and desires of those others than from supposing that they are derived from either habits (of thought) or beliefs about other features of the world (e.g. experienced regularities in the behavior of others). But the very training required to bring an animal into P&W's test situations seems to provide the relevant experience for engendering such alternate habits or beliefs. P&W are aware of this, and much of the complexity of the tests they have designed is dictated by their desire to make this alternative hypothesis less plausible. But in becoming so devious, the tests seem - to me - to sacrifice the most interesting hypothesis: it would be

much more exciting to discover that chimpanzees normally have (naturally acquire in their lives) a theory of mind then to discover that chimpanzees can have a theory of mind DINNED into them eventually. Bears can ride bicycles - a surprising fact of allusive theoretical interest. But when one tries (as I have, now, for several days) to dream up better experiments for P&W to run, one begins to appreciate that it is very hard to think up direct, natural, plausible tests. Why should this be?

Very young children watching a Punch and Judy show squeal in anticipatory delight as Punch prepares to throw the box over the cliff. Why? Because they know Punch thinks Judy is still in the box. They know better; they saw Judy escape while Punch's back was turned. We take the children's excitement as overwhelmingly good evidence that they understand the situation--they understand that Punch is acting on a mistaken belief (although they are not sophisticated enough to put it that way). Would chimpanzees exhibit similar excitement if presented with a similar bit of play acting (in a drama that spoke directly to their "interests")? I do not know, and think it would be worth finding out for if they didn't react, the hypothesis that they impute beliefs and desires to other would be dealt a severe blow, even if all the P&W tests turn out positively, just because it can be made so obvious--obvious enough for four-year-old children—that Punch believes (falsely) that Judy is in the box.

But suppose we are uncertain how to interpret the children's glee: how can we go about strengthening the hypothesis that they believe Punch believes...? We can ask them questions, particularly "why questions," but others as well ("What do you think Punch would have done if...?"). But are there nonverbal tests we can also employ? It is hard to think of any that would be decisive that wouldn't be too difficult for the children. This is because of the complexity of the "thought processes" one has to impute to any person or animal who acts on the basis of such a prediction from a theory of mind. So far as I can see, the minimally complex pattern has the following format:

1 C believes that E believes that p.

2 C believes that E desires that q.

3 C infers from his beliefs in (1) and (2) that E will therefore do x, and so, anticipating E's doing x,

4 C does y because

5 C believes that if E does x, then unless C does y, C won't get something C wants, or will get something C wants to avoid.

(This is the minimally complex pattern for doing something because you believe someone believes...: doing something in order to get someone to believe something in order to get him to do something...has a different but equally complex scenario.)

The idea experiment to establish the use of such an explanatory format will have the following features a. E's anticipated action x will be a (relatively) novel action, or at least an action that (arguably) could not be anticipated by C under the circumstances simply by virtue of being habitual for E or oft-repeated in just these circumstances. (An elegant way of accomplishing this is to ensure that the believe attributed to E in (1) is false (cl. Punch and Judy)for then E will be expected to act inappropriately to the circumstances, and hence, in all likelihood, not the way E has typically acted in

b. C's action y will also be an action as much as possible from C's natural

repertoire, rather than a highly trained artificial response, for again, arduous training procedures almost inevitably provide grief for the associationist's mind (Dennett, 1976).

C. The perceived (by C) dependence of y on x should also be natural and obvious, so that C's belief in it (5) can be attributed to C on the basis of C's straightforward observation of a relatively novel circumstance, rather than on the basis of extensive training.

Trying to design experiments to meet these conditions soon reveals the difficult. Conditions (1-3) are relatively easy to meet--one would think. For instance, suppose there is a key that E, the experimenter, uses to open the banana locker. One day two boxes, one red and one green, are placed in the scene, and C sees E put the key in the red box and leave the scene. Then C sees Sneaky Pete come in and move the key to the green box. When E returns to feed C, ex hypothesis C believes that E believes that they is still in the red box, and hence C expects E to go to the red box (since it believes that E wants to get the key).But now, how can things be rigged so there is something C might see to do that is appropriate to C's expectation (meeting conditions (4) and (5))? P&W's solution at this step is to train C to perform a sort of proto-speech-act, a prediction by choice of photograph, with the assumption that, for predictions, truth is its own reward (thus satisfying C's desire in (5)). But this is gimmicky. One would prefer to have C's action y interact more meaningfully with E's action x, but reflection reveals that this is hard to set up without resorting to another sort of gadgetry: artificial dependencies created between x-type actions and y-type actions that C might be trained to recognize.

The conclusion that seems borne in on one is that unless there is a great deal of normal interaction -- either competitive or cooperative -- between C and E, there is just no way for C to come to perceive his own actions as meaning with E's in the tight way required of step (5). One can rig it up-- e.g. C could be taught that he will get a shock if E opens the red box unless C, anticipating this, moves to a particular location -- but this requires training that removes the desired novelty listed in (a). This objection to gimmickry is not just aesthetic, of course; the more artificial the test circumstances, the more restricted the range of predictions available to the theory-of-mind hypothesis, and as noted at the outset, predictive fecundity is of the essence in this investigation.

It appears that except in tricky environments that require extensive training to produce familiarity, the only act-types that naturally meet the conditions are communicative acts, such as C's warning E, on requesting something from E, or asking E a questions; and so the problem of the training fact now pertains to the training up of communicative act types. In this regard Savage-Flumbaugh et al.'s format with Austin and Sherman looking much more straightforward and promising if the communicative mode of interaction between Austin and Sherman can be

extended to relatively novel situations without (much) additional training). But still the conclusion that would follow success in such experiments would be at best that chimpanzees can be put in complex artificial environments (artificial for chimpanzees, not for people) in which they eventually develop a theory of mind. in their natural environements there seems to be no clear need for them to develop a theory of mind about each other, and hence no compelling reason to impute it to them. But perhaps further ingenious experiments will find a way of meeting the desiderata tested and make a believe out of me.

REFERENCES

Dennett, D., International systems. Journal of Philosophy, 68:87-106. 1971.

Conditions of personhood. In: A. Rorty (ed.), The Identities of persons. Berkeley: U. Cal. Oress, 1976.

EDITORIAL NOTE

Received too late for a Response from P&W or SR&B. See Coninuing Commentary. (Ed.)