Thoughts about Thoughts: comments on Whiten and Byrne's 'Tactical Deception in Primates'

Jonathan Bennett

from: *Behavioral and Brain Sciences* 11 (1988), pp. 246–7.

Do any nonhuman animals have thoughts? If so, do any have thoughts about thoughts? At first sight it looks promising to try to get at the second question through cases of deception: we may find that one animal (Agent) is motivated by a desire to produce an erroneous or ignorant state of mind in another animal (call her Patience), which implies that Agent mentally represents Patience's state of mind to himself. Before getting into the details, let me lay out the groundwork in my own way.

In all the cases we have to consider, the upshot of Agent's conduct that is relevant to his desires is some behavior on the part of Patience. We aren't going to have evidence that he sought to alter her beliefs out of basic malice (or goodwill), wanting her to get a false (or true) belief just for its own sake. The behavior of Patience's that ministers to Agent's wants may be negative—it may consist in her not interfering, not scratching him, or the like—but that is behavior too, and I shall speak of it in the language of 'doing'. In all our cases, then, Agent does A, Patience does P, which is advantageous to Agent, and we are satisfied that this is not a mere lucky coincidence.

Two questions: **(1)** Did Agent do A intentionally, acting under the guidance of some thought of what the upshot would be? If so, then: **(2)** was Agent's intention just that Patience should do P, or did he reckon on affecting her conduct by affecting her mental state? We may be sure that the only route from his conduct to hers is through her mental state, but the question is: Was he relying on that route's being followed?

Let us start with question **(1)**. When we say that Agent did A intending to bring about result R, or because he thought that doing A would bring about R, this diagnosis is always threatened from below by the possibility that Agent did A as an instance of a drill, a pattern of stimulus and response: Agent acts in circumstances of physical kind K_c, and A is of a physical kind K_a, and Agent has found that whenever he is in K_c circumstances he performs a K_a action R happens. If that is the case, Agent may have the thoughtlessly mechanical habit of performing a K_a in K_c circumstances whenever he wants R. How can this challenge from below he fended off?

One might answer: 'Well, if Agent's action A and circumstances C do not belong to any kinds K and C such that he has found in the past that in C circumstances A actions lead to R, his doing A on this occasion can't be something he does as a matter of a drill that has been inculcated in him by his past experience.' A few decades ago, that answer was given

by psychologists who thought they had a viable concept of animal 'insight' that could be explained in terms of radically unprecedented behavior. But that was all a muddle. If the connection between A and R is not somehow attested to in Agent's past experience, his doing A in order to get R on the present occasion becomes not insightful but merely lucky or else miraculous. For a post-mortem on the 'insight' muddle see Bennett, *Rationality* (1964).

The right way to meet the challenge from below is not to find behavior that doesn't instantiate a pattern, but rather to find behavior that falls into a teleological pattern and into no one stimulus-response pattern: that is, a kind of result that Agent often brings about by movements of many different kinds, on the basis of many physically different clues that the result is achievable. This approach takes us away from 'when Agent gets sensory input from a K_c environment he makes movements of physical kind K_a toward something more like 'when Agent has evidence that R can be achieved he does whatever will produce R'. Of course it's much more complicated than that, but that outlines what is chiefly needed.

So the conclusion that Agent is acting intentionally—that is, behaving as he does because of what he thinks and wants—does not conflict with the need for pattern, regularity, repetition, so long as the patterns are not stimulus-response ones but rather are teleological in the way I have explained.

Now, suppose we are satisfied that much of Agent's behavior is intentional, including some in which he intends to modify the behavior of Patience. We want to know whether his belief that by doing A he will get Patience to do P is ever based on his belief that by doing A he will affect her mental state in a certain way.

The evidence that Agent is a 'psychologist', as Whiten & Byrne (W&B) put it, goes like this: Agent believes something of the form: 'If I do A, Patience will do P', and we want to know why he connects his doing A with her doing P. If we can't explain this better, that is, more economically, than by crediting him with believing **(1)** that if he does A she will go into mental state M, and **(2)** that if she goes into mental state M she will do P, then we have a case for attributing those two beliefs to Agent and thus crediting him with thoughts about Patience's mental state.

To be fully entitled to attribute beliefs **(1)** and **(2)**, we would need evidence that Agent has had opportunities to learn that those two are true. That is a complex matter I don't fully understand; to sort it out, we would need to understand how Agent's experience of his own mind relates to his beliefs about other minds. I shall restrict myself to the more immediate question of challenges from below—that is, of what would undermine the attribution to Agent of beliefs **(1)** and **(2)** even if there were no problems about learning.

The immediate threat is that Agent can be understood to have connected his doing A with Patience's doing P in some manner that doesn't run through Patience's psyche. That will be the case if A is of some physical kind K_A, and P is of a physical kind K_p, such that Agent's experience has accustomed him to its being the case that when he does something of kind K_A Patience does something of kind K_p. If K_p really is a physical kind, and doesn't have to be marked out in terms of psychological underlay ('movement that indicates her lack of interest', 'movement that she wouldn't make if she were afraid'), Patience's mind is banished from Agent's scenario and the challenge from below has succeeded.

From this I conclude that most of the anecdotes W&B have collected are at best weak evidence that Agent is a psychologist.

The 'hiding from view' cases are impressive only to the extent that in them Agent undergoes some quite complex maneuvering to keep something out of Patience's view: That is indeed evidence of 'intentionalness', acting toward a foreseen outcome. However, it doesn't constitute evidence that Agent has thoughts about Patience's mental state unless there is pressure to suppose that the outcome, as represented in Agent's mind, is some state of Patience' s mind. That pressure is weak. It seems possible, even plausible, to suppose that many animals at various levels have a *physicalistic* notion of line of sight, based on proximity and absence of intervening objects. Agent's grasp of the advantages of keeping something out of Patience's line of sight probably doesn't require him to operate as a psychologist any more than does his operating to keep downwind of his prey.

Those auditory examples in which the deceptive behavior consists in keeping quiet are even weaker as evidence of thoughts about mental states. Agent needs only to connect his silence with Patient's noninterference, and that he can presumably do by simple induction. *Keeping quiet* is an intrinsic, physical kind of behavior; it lacks the complexity of some of the visual examples, and is therefore less good as evidence that these cases involve intentionalness at all, let alone intentions to produce false beliefs. *Not interfering* is not intrinsic, because it means 'not behaving in a manner that stops me from getting what I want'; but that doesn't help much. It doesn't even seem to involve Agent's thinking about Patience's thoughts, and it's not especially impressive in any other way. Plenty of fairly low-level nonhuman behavior can't be understood unless the animal can recognize external events as threatening, unwelcome, interfering, and the like. A sense of how external events relate—whether as conducive or threatening—to one's own desires is required for any kind of cognitive mentality.

In those cases, then, all Agent needs to have learned is that in certain familiar kinds of situations his silence is a means to Patience's noninterference; and there is really nothing left of the case for thinking that Agent is a psychologist in these situations.

Similarly with distraction by looking away: Agent needs only to know that he and his kind tend to look in directions in which others look, and don't continue with attacks when they are looking off in another direction. That challenge from below presupposes that Agent has a grasp of 'looking in direction D' as a physical kind of behavior, marked off by posture, direction in which eyes are pointing, eyes open, and so forth, and not in terms of anything mentalistic. This—which could also be used to amplify the line-of-sight notion mentioned above—seems to be a modest assumption that is well supported by the data. **(**What it may imply is: Agent knows that in his community when one looks in a particular direction, so do others who see him do so; this knowledge is a conjunction of two bits of information: one about what happens when *he* looks in a given direction, and the other about what happens when *others* do so. If his thought about where others are looking is essentially a thought about posture and such, then we mustn't assume that he can simply generalize from the consequences of their looking in a given direction to the consequences of his doing so. Whether that is so depends on how Agent's experience of his own body relates to his perceptions of the bodies of others.**)**

Those remarks apply, *mutatis mutandis*, to the 'inhibition of attending' cases as well. W&B themselves notice the structural similarity between these two kinds of cases.

The reported cases of distraction-by-leading-away don't create any case for regarding Agent as a psychologist, so far as I can see. Even if the leading away is deliberate, and is

intended to get Patience out of the way, it could be based on a grasp of 'Where I go [smacking my lips, etc.], she goes', with no thought about Patience's state of mind. W&B point to the possibility that Agent is sensitive to Patience's level of attention: If she is not attending to him, she won't be drawn away by him; if she continues attending to him, she will return when he does. There could be (though I gather that there isn't yet) evidence that Agent's behavior in this general category reflects a sensitivity to those differences in Patience; That might add to the plausibility of a 'thought about thought' diagnosis, but it might not. It would fail if the relevant notion of attending could be well understood in physicalistic terms, along the lines of my suggested account of 'looking in direction D'.

Similar remarks apply to the cases of distraction through intimate behavior, and I think they can be extended to the various kinds of deception W&B present in sections 2.5.3–2.5.5.

Having described Whiten & Byrne's problem situation in my own way, and expressed doubts about how far they have got with it, I want to add that I admire their grasp of what their problems are and of what would solve them. This is a useful and interesting paper, and I am glad to have a chance to try to push it further in its right direction.