© Blackwell Publishers Ltd. 1999, 108 Cowley Road, Oxford, OX4 1JF, UK
and 350 Main Street, Malden, MA 02142, USA.
Mind & Language, ISSN: 0268-1064
Vol. 14. No. 1 March 1999, pp 23–31.

What Autism May Tell Us About Self-Awareness: A Commentary on Frith and Happé

DIANA RAFFMAN

For philosophers of mind, what is especially intriguing about Frith and Happé's discussion is the idea that autistics know their own minds only, or almost only, by observing their own behaviour. The authors do not say exactly how the Asperger subjects gain this knowledge of themselves; the clearest statement I can find occurs on p.11: 'If the person with autism can judge their own inner states *only by their actions*, it might be important to teach behaviours which express for oneself what one thinks and feels' (emphasis added). Nevertheless, I take Frith and Happé's central hypothesis to be that instead of having the (apparently) direct or immediate or non-inferential self-knowledge we normals take for granted, autistics know what is transpiring in their own minds only by applying an explicit theory of mind to their own behaviour.[1] They do not know their own minds in the normal way, by introspection.

The asymmetry between first-person introspective knowledge and third-person inferential knowledge of the mental states of others has long been a topic of interest to philosophers. As Paul Boghossian explains,

> [there appears to be] a profound asymmetry between the way in which I know my own thoughts and the way in which I may know the thoughts of others. The difference turns not on the epistemic status of the respective beliefs [e.g., not on any infallibility of my self-knowledge—DR], but on the manner in which they are arrived at, or justified. In the case of others, I have no choice but to *infer* what they think from observations about what they do or say. In my own case, by contrast, inference is neither required nor relevant. Normally, I know what I think . . . without appeal to supplementary

I am grateful to Kathleen O'Dowd, George Pappas, and the editors of *Mind and Language* for helpful commentary.
Address for correspondence: Diana Raffman, Department of Philosophy, Ohio State University, 350 University Hall, 230 North Oval Mall, Columbus, OH 43210–1365, USA.
Email: Raffman.1@osu.edu.

[1] A limited vindication of Gilbert Ryle, it would seem; see Ryle, 1949.

> evidence. Even where such evidence is available, I do not consult it. I know what I think directly. I do not defend my self-attributions; nor does it normally make sense to ask me to do so.[2]

The nature and etiology of our knowledge of others' minds seems clear enough: we observe their behaviour and infer from that evidence what they think. But the immediacy of our knowledge of our own minds is mysterious: it seems that we *just know* what we are thinking, that we don't need to collect *evidence*. If Frith and Happé are right, Asperger subjects lack this immediate rapport with their own thoughts; they do need to collect (behavioural) evidence in order to know what they themselves are thinking. Following a long and fruitful tradition of studying abnormal psychological processes in order to understand normal ones, it may be that if we can figure out what goes wrong in the autistics, we can shed some light on the nature of self-knowledge in normal subjects. That would be exciting indeed.

My role as a commentator, of course, is to raise questions for further reflection. To that end, let me introduce a distinction between two interpretations of the claim that an individual fails to be self-aware, fails to know or 'have access' to the contents of his own mind by introspection. (It is surely not an exhaustive distinction.) On the one hand, such a claim might mean that an individual lacks the concepts of belief and desire and the rest in terms of which we ordinarily conceive and express our mental states and explain and predict our own behaviour and the behaviour of others. In other words, the claim might mean that an individual is not competent with what Frith and Happé call the Theory of Mind ('ToM'). For all this first interpretation says, if such an individual were to acquire the relevant ToM concepts, he could employ them to express self-knowledge gained in the normal direct way. On the other hand, an individual might lack self-knowledge, might be unaware of himself, in a second, stronger sense. In this sense, even if he mastered the ToM concepts, still he could apply them to his own case only by observing his own behaviour.

The interesting idea advanced by Frith and Happé is that the Asperger subjects are unaware of themselves in this second, stronger sense: even when they have mastered the relevant ToM concepts, they can apply those concepts to their own case only indirectly, by inference from observation of their own behaviour. What I will suggest, however, is that the evidence Frith and Happé present is compatible with the hypothesis that the Aspergers subjects lack self-knowledge only in the first, weaker sense, and so does not yet justify the more extravagant hypothesis the authors endorse. Call the first, weaker hypothesis the 'Conceptual Incompetence' hypothesis or 'CI' for short. Strictly speaking, we ought to distinguish at least three hypotheses one might advance about the Asperger subjects: (1) they fail to have ('first

---

[2] Boghossian 1989, p. 7. As will become clear, I draw much from this elegant paper in my discussion here.

© *Blackwell Publishers Ltd. 1999*

order') conscious mental states to be aware of; (2) they have conscious mental states but are not aware of them; and (3) they have conscious mental states and are aware of them but lack the conceptual competence to represent and report them in ToM terms, i.e. to represent and report them *as* beliefs and desires and the rest. I will suggest that the evidence provided by Frith and Happé is compatible with hypothesis (3), *viz*. CI, and so fails to justify the more extravagant hypothesis (2). At the end I will also suggest that hypothesis (1) applies in some cases—not so much that autistics sometimes have *no* conscious mental states but rather that they sometimes have bizarre and/or impoverished conscious states; hence it should come as no surprise that their self-ascriptions of mental states are correspondingly bizarre in such cases.

In order to defend CI (3) I will need first to clarify and defend the idea of a species of self-awareness (self-consciousness, self-reflection) that does not involve the mobilization of ToM concepts. Progress here is complicated by the many different conceptions of self-awareness in the philosophical literature and elsewhere. Self-awareness may or may not involve the mobilization of concepts; it may or may not be tied essentially to the ability to make verbal reports; it may or may not involve attentional mechanisms; it may or may not be a so-called cognitive achievement (more on this in a moment); and so on. My goal here is not to attack any of these conceptions, but rather to defend one that has itself come under attack (see just for example Davidson, 1987 and Burge, 1988). To do this, I will need to say a little about the history of philosophical views of self-knowledge.

According to philosophic tradition, inherited largely from Descartes, the mind is both unified and transparent to itself; that is, each of us has complete and infallible knowledge of the contents of his own mind. If I am in pain, I know that I am, and if it seems to me that I'm in pain, then I am in pain; if I believe it's raining outside, then I know I believe it, and if it seems to me that I believe it, then I do believe it; similarly for other mental states. There is, as it were, no distinction between appearance and reality in respect of self-knowledge. Contemporary philosophy of mind, by contrast, discards much if not most of this traditional picture. Science has taught us that the contents of our minds are largely hidden from us and that we are frequently mistaken about what goes on in there. Evidently we harbour all manner of tacit knowledge and unconscious mental activities whose presence may be more obvious to others than to ourselves and which in any case we must infer from observation of our own behaviour. To be sure, certain elements of the Cartesian view are retained—in particular the idea that we occupy a privileged epistemic position in respect of the contents of our own *conscious occurrent* states (perhaps especially our occurrent sensations). The assumption is that we normally have better access to our own conscious occurrent states than others do. (This idea is often expressed by saying that we enjoy *first person authority* with respect to these states.) Nevertheless, the possibility of error is always open.

We need to distinguish at least two kinds of errors one might make about

© *Blackwell Publishers Ltd. 1999*

the contents of one's own mental states. On the one hand, one might misidentify a state because it occurred in the past, or because it is unconscious—because the state is not currently available, as we might say. On the other hand, one might misidentify even an occurrent conscious state in the sense of judging it to be a pain as opposed to an itch, say, or a fear as opposed to a hatred. Paul Churchland illustrates:

> Consider the occurrence of something rather *similar* to pain—a sudden sensation of extreme cold, for example—in a situation where one strongly *expects* to feel pain. Suppose you are a captured spy, being interrogated at length with the repeated help of a hot iron pressed briefly to your back. If, on the twentieth trial, an *ice cube* is covertly pressed against your back, your immediate reaction will differ little or none from your first nineteen reactions. You almost certainly would think, for a brief moment, that you were feeling pain . . . [Your] judgment would be false. (Churchland 1984, p. 77)

One's access to one's occurrent conscious states is by definition immune from the first sort of error I just described (*viz*. an error about past or unconscious states), but it is not immune from the second: identifying a state as a pain or as a fear involves the application of the ToM concepts of pain and fear, and even assuming that one is competent in the use of these concepts, it is always possible to misapply them in a given case.

For my purposes the important point here is this. The possibility of mistakes of this latter kind, namely misidentications of one's own occurrent conscious states, suggests that self-knowledge of such states is what some philosophers have called a *cognitive achievement* (see e.g. Boghossian 1989, pp. 19–20): it is *in some sense based on evidence*. Boghossian observes that

> the difference between getting it right and failing to do so (either through ignorance or through error) is the difference between being in an epistemically favorable position with respect to the subject matter in question—being in a position to garner the relevant evidence—and not. To put this point another way, it is only if we understand self-knowledge to be a cognitive achievement that we have any prospect of explaining its admitted shortcomings.[3]

Whether what I am referring to as 'evidence' is properly so-called is a matter of dispute (some philosophers, particularly those who favor an externalist epistemology, will dislike this use of the term); but I think we can avoid

[3] 1989, p. 19. In further support of the idea that self-knowledge is a cognitive achievement, Boghossian points out (1) that 'self-knowledge can be directed: one can decide how much attention to direct to one's thoughts or images, just as one can decide how much attention to pay to objects in one's visual field'; and (2) that self-knowledge is 'subject to cultivation or neglect'.

© *Blackwell Publishers Ltd. 1999*

those controversies here. What I have in mind is just that while my introspective judgements are not based on behavioural evidence, still there is something—something that is in some sense consciously accessible to me—in virtue of which my introspective judgements are correct or not. (In the case of a sensory state like pain, we can think of this something as, well, the *phenomenology* itself, the *experience* I am now having.) I will use the term 'introspective evidence' to refer to this something that makes my introspective judgements right or wrong.

N.B. It must be acknowledged that the existence and nature of such introspective evidence, the 'something' upon which the application of ToM concepts to one's own case is based, is a difficult and highly controversial issue in philosophy of mind; and I have no developed theory of it to offer. Nevertheless, I think we may fairly appeal to the notion here, given the plausibility of viewing self-knowledge, even self-knowledge of occurrent conscious states, as a cognitive achievement and, as we shall soon see, given that it seems to make the best sense of the Asperger subjects Frith and Happé describe.

This view of self-knowledge as a cognitive achievement would seem to make room for a distinction between two senses of the term 'self-awareness'. The thought is that in one sense, self-awareness involves the application of ToM concepts in a judgement; here, self-awareness is properly regarded as self-*knowledge*. I will reserve the term 'introspection' to refer to the application of ToM concepts in judgements of this kind. In another, thinner sense, self-awareness is a non-conceptual (or anyway non-ToM-conceptual) awareness of one's conscious mental states, perhaps consisting in or otherwise involving the direction of attention upon those states. In the terminology we have adopted here, self-awareness in this second sense is the source of the introspective evidence upon which self-awareness in the first sense—self-knowledge—is based. (Self-awareness in the second, thinner sense may lie at the heart of our persistent intuitions of infallibility: insofar as we make no judgement, apply no concepts to the states of which we are self-aware, we cannot be mistaken about them.) If something like this view is right, it may be that our self-knowledge seems direct, noninferential, immediate because our practice of self-applying the ToM concepts is so ancient and so practiced that it proceeds automatically. We normals have become experts with ToM. Indeed, the ToM framework may be so deeply ingrained that our conscious states seem to present themselves to us *ab initio*, as it were, *as* beliefs and desires.[4]

What I want to suggest, then, is that the descriptions and testimony of the Asperger subjects presented by Frith and Happé are compatible with the hypothesis that these subjects are self-aware in the second sense but not in the first. They are aware of their own states—they have the relevant introspective evidence—but lack the competence with ToM concepts required to

[4] I thank George Pappas for discussion of this point; see also Pappas, 1996.

© *Blackwell Publishers Ltd. 1999*

conceive and report those states (that evidence) in ToM terms. On this view (CI), the reason for the autistic's 'effortful' self-ascriptions of mental states is not the difficulty of inferring what he thinks from his observations of his own behaviour, but rather a difficulty in mastering (understanding?) the ToM conceptual framework itself, or, perhaps more likely, a difficulty in learning to apply the ToM concepts to the conscious states of which he is aware. Such a view sits well with the fact that the ability to ascribe ToM mental states to others appears to go hand in hand with the ability to ascribe ToM mental states to oneself. If both abilities employ the same conceptual framework, namely ToM, conceptual incompetence in the one case would *eo ipso* make for conceptual incompetence in the other. Here one is reminded of the insights of P. F. Strawson:

> There is no sense in the idea of ascribing states of consciousness to oneself, or at all, unless the ascriber already knows how to ascribe at least some states of consciousness to others (p. 106). . . . [I]t is essential to the character of these predicates that they have both first- and third-person ascriptive uses, that they are both self-ascribable otherwise than on the basis of observation of the behaviour of the subject of them, and other-ascribable on the basis of behaviour criteria. To learn their use is to learn both aspects of their use. In order to *have* this type of concept, one must be both a self-ascriber and an other-ascriber. In order to *understand* this type of concept, one must acknowledge that there is a kind of predicate which is unambiguously and adequately ascribable *both* on the basis of observation of the subject of the predicate *and* not on this basis (p. 108) . . . (Strawson 1959)

Consider now Frith and Happé's discussion of the Asperger subjects. Two sets of considerations in particular seem to me to favor CI over Frith and Happé's stronger hypothesis (my (ii)). First are the autobiographical reminiscences of early childhood recorded by Williams (1994), Jolliffe, Lansdown, and Robinson (1992), Grandin (1984), and Gerland (1997). For example: 'I was sick to death of my attention wandering onto the reflection of every element of light and colour . . .' (Williams); 'I sometimes got annoyed once I realized that I was expected to attend to what other people were saying' (Jolliffe, Lansdown, and Robinson); '[t]he feeling was like a constant feeling of stage fright all the time' (Grandin); '[t]ime and again I was very hurt when people said they knew things about me . . .' (Gerland). These Asperger subjects are currently reporting their past mental states—states they underwent, we are to assume, before they acquired the ability to conceive and report them *as* wandering attention, annoyance, fright, hurt, and so on. If Frith and Happé's hypothesis is correct, these reports are based upon the Aspergers' observations (or recollections) of their own behavior. It is difficult to see how this could be so, however. While some of their reminiscences do concern past behaviours (e.g., 'As a child, I often talked out loud . . .'; 'I

© *Blackwell Publishers Ltd. 1999*

remember minutely observing how the sand flowed . . .'), it seems unlikely that the Aspergers can remember their past behaviour generally with the degree of detail required for the reminiscences.

It also does not seem plausible that the Asperger subjects could draw the requisite inferences from their *current* behaviour. Presumably the idea here would be that the Asperger subjects (a) have some form of memory of their past experiences and feelings though they were never directly aware of them (i.e., never introspected upon them), (b) currently behave in a manner responsive to those remembered experiences and feelings (perhaps in a kind of reenactment of past behaviours?), and then (c) infer from observation of their current behaviour that they had those experiences and feelings in the past. I cannot prove that such a view is incorrect, but it seems implausible in the extreme, and in any case Frith and Happé supply no evidence to suggest that such current behavioural responses to remembered experiences goes on. In addition, since the ability to self-ascribe mental states was acquired by the Asperger subjects at a time later than the time of the remembered experiences, they can't be making their current self-ascriptions by remembering past self-ascriptions, i.e. by remembering past verbal behaviours.

Rather, it seems more plausible to suppose, with CI, that these subjects were indeed aware of their experiences and feelings in the past—they possessed introspective evidence of the thin kind I have tried to clarify here—but lacked the ToM concepts needed to conceive and express them in the normal way. Indeed, Frith and Happé themselves write that, apart from 'effortful' practice at making ToM ascriptions, the autistics 'lack the cognitive machinery to represent their thoughts and feelings *as* thoughts and feelings' (pp. 7–8), and that some of the experimental 'results suggest that the children with autism did not conceptualize their own intentions as intentions' (p. 9). At the very least, CI is *compatible* with the character of the Aspergers' reminiscences, and so Frith and Happé's stronger hypothesis (2) is not yet warranted.

Second, Frith and Happé cite evidence suggesting that the ability to ascribe mental states to oneself and the ability to ascribe them to other people, in normals and autistics alike, are underwritten by a common cognitive mechanism; among other things, 'there is little evidence from the developmental literature to suggest that mental states are attributed to self before they are attributed to others' (p. 5). (The apparent selective impairment of the autistics' ability to ascribe ToM states both to themselves and to others tends to support the further hypothesis that this common mechanism is innate or 'wired in'.) As Frith and Happé themselves acknowledge, however, the idea of a common mechanism is *prima facie* counterintuitive: for the kinds of reasons I discussed above, intuition suggests that awareness of one's own mental states and awareness of others' mental states employ radically different mechanisms. CI, in contrast, allows that self-knowledge and knowledge of other minds are subserved by different (though still possibly innate) mechanisms in both autistics and normals. A proponent of CI can say that

© *Blackwell Publishers Ltd. 1999*

what self-knowledge and knowledge of other minds have in common is not their etiologies or manners of acquisition (the 'manner in which they were arrived at', as Boghossian puts it), but rather the conceptual framework in terms of which they are formulated and made usable for explaining and predicting behaviour. If there is a common mechanism or module, it is more likely to underlie that conceptual competence and the consequent ability to *ascribe* mental states to oneself and others.

CI also appears compatible with other evidence presented by Frith and Happé. For instance, it sits well with the fact that Asperger subjects acquire the ability to ascribe mental states to themselves and others only with considerable effort and explicit training. Many activities, including the use of language and mathematical calculation, can be mastered only laboriously when mastered late, and may never be executed with the ease of one who learns them in early childhood. Competence with the ToM framework may well be the same. (Indeed, perhaps there is a 'critical period' for the development of ToM competence; results from the so-called false belief tasks cited by Frith and Happé may support this idea.)

Finally, it is worth noting that much of the evidence Frith and Happé present to show impaired self-awareness or self-knowledge in autistics seems equally interpretable as showing only impaired (i.e., unusual or even bizarre) *experience*. The discussion of 'abnormal sensory and pain experiences', 'hypo- and hyper-sensitivity to sound, light or touch', and much if not most of the subject testimony cited by Frith and Happé is interpretable in this way, it seems to me. So the authors will need to make clear how the evidence they cite pertains to the 'second order' states as opposed to the 'first order' ones. Whatever the results of that further work, the evidence Frith and Happé present, and their discussion of it, will be of very significant interest to philosophers of mind.[5]

Department of Philosophy
Ohio State University

References

Boghossian, P. 1989: Content and Self-Knowledge. *Philosophical Topics*, 17, 5–26.
Burge, T. 1988: Individualism and Self-Knowledge. *Journal of Philosophy*, 85, 649–63.

[5] I should also point out that evidence from autistic subjects is often cited in the controversy between the so-called *theory theory* and the so-called *simulation theory* of the ability to ascribe mental states to others. Perhaps not surprisingly, proponents of both views cite the autistics as tending to confirm their respective positions. As far as I can see, the evidence from autism is neutral; it favors neither position over the other. In particular, even if the Asperger subjects are best understood as applying to (themselves and) others an explicit *theory* of mind, this shows nothing about what goes on in the normal case; among other things, it may be that the autistics must resort to use of a theory because they lack the normal ability to simulate other people. The interested reader will want to have a look at Davies and Stone, 1995 for a helpful introduction to these issues.

© *Blackwell Publishers Ltd. 1999*

Churchland, P. 1984: *Matter and Consciousness*. Cambridge, MA: MIT Press.
Davidson, D. 1987: Knowing One's Own Mind. *Proceedings of the American Philosophical Association*, 60, 441–58.
Davies, M. and Stone, T. (eds) 1995: *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell.
Gerland, G. 1997: *A Real Person: Life on the Outside*, translated from the Swedish by J. Tate. London: Souvenir Press.
Grandin, T. 1984: My Experiences as an Autistic Child and Review of Selected Literature. *Journal of Orthomolecular Psychiatry*, 13, 144–75.
Jolliffe, T., Lansdown, R., and Robinson, C. 1992: Autism: A Personal Account, *Communication*, 26, 12–9.
Pappas, G. 1996: Experts, Knowledge, and Perception. In J.L. Kvanvig (ed.), *Warrant in Contemporary Epistemology*. Lanham: Rowman and Littlefield, 239–50.
Ryle, G. 1949: *The Concept of Mind*. London: Hutchinson.
Strawson, P. F. 1959: *Individuals: An Essay in Descriptive Metaphysics*. London: Methuen.
Williams, D. 1994: *Somebody Somewhere*. London: Doubleday.

© *Blackwell Publishers Ltd. 1999*