Philosophical Books Vol. 50 No. 4 October 2009 pp. 227–234

CRITICAL NOTICE OF LUCY O'BRIEN, *SELF-KNOWING AGENTS*[1]

JOHANNES ROESSLER
University of Warwick, United Kingdom

What is the relation between first-person thought and self-consciousness? On one view, grasp of the first person is based on acquaintance with the self. Self-consciousness makes first-person reference possible, in much the same way in which perceptual acquaintance with objects of 'outer sense' makes perceptual demonstratives such as 'this lemon' available to us. The 'perceptual demonstrative model' of first-person reference is widely regarded as discredited though. It is not just that the task of demystifying the idea of acquaintance with the self presents formidable challenges. Nor is the problem just that the 'perceptual demonstrative model' is hard to reconcile with certain basic features of first-person thought, such as its immunity to reference failure. The fact is that in contrast to perceptual demonstratives, the first person is a token reflexive. The reference of any serious use of 'I' is determined by a completely general rule, the self-reference rule (SRR): 'I' refers to whoever uses it. It is not determined by acquaintance with the object referred to. Grasp of the first person is a matter of being able to use an expression or concept governed by SRR.

Lucy O'Brien's *Self-Knowing Agents*[2] offers, among many other things, a rich and subtle critique of what she calls the 'self-reference approach' to the first person. O'Brien acknowledges that the approach marks an important advance over the 'perceptual demonstrative model.' Her dissatisfaction with it stems from its perceived failure to address, let alone answer, what she thinks of as the deepest and most important question in this area: the question of how to understand the relation between grasp of the first person and self-consciousness. O'Brien calls her proposed alternative to, or supplementation of, the self-reference approach the 'agency account.' The account is intended to illustrate and substantiate the general thesis that activity, both psychological and physical, 'has an important role to play in addressing some of the central problems that subjectivity presents us with' (vii). In O'Brien's view, philosophers grappling with the problems of subjectivity have tended to make these difficult problems more difficult still by paying insufficient attention to the fact that we are active creatures. The first part of the book looks at the specific problem of first-person reference. The second part contains additional material on the nature and first-person epistemology of bodily action and on bodily awareness. Here I will focus on O'Brien's treatment of the first person. Her distinctive emphasis on the connection between the first person and self-consciousness

1. An earlier version of this discussion was presented at a symposium on *Self-Knowing Agents* at the meeting of the APA (Pacific Division) in Vancouver in April 2009. I'd like to thank Lucy O'Brien for extremely helpful comments.
2. Lucy O'Brien, *Self-Knowing Agents* (Oxford University Press, 2007, 231 pp).

seems to me convincing, original, and important. My question is whether O'Brien is right that the 'self-reference approach' is not in a position to explain and respect that connection, and that the 'agency account' is.

1. The Agency Account

O'Brien writes:

> In essence, the problem is that any attempt to explain first-person reference as 'reflexive' reference runs into trouble, because reflexive reference can only be first-person reference if one knows that one is referring to oneself. However, that knowledge then also needs explication. It can seem obvious however that knowing that one is referring to oneself involves referring to oneself first-personally. But if that is so it seems one cannot give a non-circular account of first-person reference. (p. 8)

As O'Brien presents the dialectical position, there are two ways to deal with this problem. One option is to reject the assumption responsible for it, that first-person reference involves knowledge that one is referring to oneself (or, as she sometimes puts it, that the first person 'expresses self-consciousness'). It might be argued that first-person thought is to be found in relatively unsophisticated creatures—creatures to whom *we* may wish to attribute first-person beliefs, for example about their own spatial location, but who do not themselves attribute first-person (or for that matter, any other) beliefs to anyone and who do not have the concept of meaning or reference. O'Brien raises some interesting objections to this 'reductionist approach,' as she calls it. I will not consider this here. The remaining option is what she calls the 'two-tier strategy,' according to which self-conscious reflexive reference is to be understood as the joint upshot of two factors: our grasp of SRR plus some independent capacity for first-person reference or self-consciousness. The point of introducing this latter element is that it is expected to enable us to provide the required 'explication' of a first-person user's knowledge that she is referring to herself. Just to illustrate, a crude version of the two-tier strategy would hold that we need to distinguish two kinds of first-person concept: the token-reflexive, governed by SRR, and an introspective demonstrative, whose reference is determined not by SRR but by appeal to the relation of 'introspective acquaintance' that makes the demonstrative available to us. One's grasp of the former draws on one's grasp of the latter. When using the token-reflexive, one knows who one is referring to by dint of one's knowledge of SRR and one's introspective awareness that it is 'this [introspectively presented] thinker' who is producing the relevant token of 'I.' The second tier of the strategy enables us to give a noncircular 'explication' of one's knowledge that one is referring to oneself.

This picture is of course completely incredible. O'Brien concurs with this judgment and indeed makes a compelling case for it in her discussion of the perceptual demonstrative model. She also criticizes a less crude version of the 'two-tier strategy,' which she finds in Peacocke's *Sense and Content*. Her

reaction to the apparent failure of the 'two-tier strategy,' though, is not to abandon the strategy but to keep on refining it. A step in the right direction, she thinks, would be to make the second tier nonconceptual—to invoke not a foundational first-person *concept*, but a 'capacity for first-person reference at the non-conceptual level' (p. 71). The trouble with a capacity for first-person *reference*, however, is that it still raises the very questions it was intended to answer: of how the reference of the nonconceptual representation is determined and—supposing the answer to this question appeals to something like SRR—the question of how it is that the representation in question 'expresses the subject's self-consciousness' (p. 71).

To overcome this final hurdle, O'Brien suggests, 'we must identify a form of self-consciousness that is prior to, and independent of, our capacity for first-person reference.' (p. 73). There is a faint echo of Schopenhauer in O'Brien's fascinating proposal at this point: the answer to the riddle is *will*—or at least, a sense of control. O'Brien argues that when we are engaged in activities, mental or physical, we exercise rational control, and such control involves a distinctive sort of experience. There are two different constructions she uses in talking about this experience. She sometimes simply refers to conscious actions ('the act produced by a process of considering what to do is a conscious act') (p. 119); more often, she speaks of an 'agent's awareness' *of* her action ('I hold that agents' awareness of an action is a product of its being brought about by a certain process of rational assessment which results immediately in action.') (p. 121). Importantly, 'agent's awareness' must not be confused with propositional knowledge that one is acting. What O'Brien has in mind is a type of experience she thinks constitutes the *source* of such knowledge. The background assumption here is that 'there is a general entitlement immediately to self-ascribe those states and activities (. . .) which are conscious' (p. 119). Her central idea is that 'agent's awareness' provides a (pre-first-personal) *basis* for knowledge that one is referring to oneself and thereby provides the resources we need safely to cross the minefield that is the 'two-tier strategy.'

O'Brien's theory is obviously immensely complex and ambitious. A proper assessment of her theory would need to look in detail, first, at her analysis of the phenomenology of action, second, at her account of the epistemic role of 'agent's awareness,' and third, at the way she puts both sets of claims to work in accounting for first-person reference. I will concentrate on this third level of her theory, though one of the two questions I want to raise has obvious reverberations across the other levels. My first question is: why should we accept the noncircularity requirement that drives the 'agency account?' My second question is: is 'agent's awareness,' as O'Brien conceives it, a form of *self*-awareness?

2. The Self-Reference Approach

Suppose it occurs to you 'it's time I went down for dinner.' Such episodes of thought are normally open to reflective knowledge—perhaps broadly of the type 'knowledge of what one is doing.' For example, you may reflect that you

think that it's time you went down for dinner, and further, that in thinking this, you are having a thought about yourself, a thought in which you are referring to yourself. All this, of course, involves the use of the first-person as governed by SRR, both in the content of the self-ascribed thought and in self-ascribing it. Why should a self-reference theorist be troubled by this fact? Why should we demand a noncircular 'explication' of your realization that you are the subject and object of your thought?

According to O'Brien, the problem with the self-reference approach is that it cannot provide a satisfactory analysis of what it is to think of oneself first-personally. For she thinks that systematically reflexive reference—reference in accordance with SRR—is not sufficient for *first-person* reference.

In other words, the self-reference theorist is presented as holding the following:

> (1) Anyone who uses a term or concept governed by SRR will refer to themselves first-personally. (p. 7)

However, (1), O'Brien argues, is open to counterexamples: cases in which an expression is used in accordance with SRR by creatures that lack self-consciousness (and hence, are incapable of first-person thought). So SRR is not sufficient to (as she puts it) 'individuate' first-person reference. It is at this point that circularity becomes an issue. Suppose we amend (1) by adding the condition that the subject realize or know that she is referring to herself, where this realization involves the use of the first person. As an analysis of what it is to think of oneself first-personally, the amended account looks objectionably circular. The aim of the 'agency account' is to provide a noncircular analysis of first-person reference. It claims that using a referential device in accordance with SRR *plus* enjoying 'agent's awareness' of doing so are jointly sufficient conditions for first-person reference.

Sometimes, O'Brien gives the impression that in addition to the 'individuation problem,' the self-reference approach also faces an independent 'explanation problem,' the problem of explaining 'how it is that "I" expresses self-consciousness' (p. 57). However, it seems to me she offers no grounds, independently of the individuation problem, for thinking that the required explanation has to satisfy the condition of 'noncircularity.' As far as I can see, the motivation for the agency account hinges on the alleged insufficiency of (1). Actually though, it turns out to be quite difficult to produce credible counterexamples to (1). Of course, as Anscombe emphasized, we need to distinguish between self-reference and first-person reference. Oedipus was referring to himself when talking about the slayer of Laios, but he did not entertain a first-person thought. But then 'the slayer of Laios' is not governed by SRR. O'Brien briefly considers Anscombe's tale of a society of speakers who use 'A' (a name stamped on the inside of their wrists) to refer to themselves but who lack self-consciousness. However, O'Brien immediately (and rightly) sets this to one side since once again, 'A,' in Anscombe's story, is not a device whose reference is determined by SRR. It is possible to think

of cases where 'A' is used to refer to someone other than oneself.[3] O'Brien's strongest argument features David Lewis' example of two gods. One god lives on the tallest mountain and throws down manna. The other lives on the coldest mountain and throws down thunderbolts. Both gods are omniscient as far as descriptive, nonindexical knowledge is concerned. However, neither of them knows which of the two gods he is; neither of them has self-consciousness. In O'Brien's version of the story, the two gods do use the first person, and they grasp SRR. Thus, if the god on the tallest mountain utters 'I am throwing down manna,' both gods know that 'I' in this utterance refers to the god on the tallest mountain. There is, in some sense, systematically reflexive reference here, but no self-consciousness, hence no first-person reference.

One response to this argument is to restrict the scope of (1) to *rational thinkers*:

> (1') Any rational thinker who uses a term or concept governed by SRR will refer to themselves first-personally.

It is a time-honored idea that there is a deep connection between rationality and first-person thought. As McDowell put it in a discussion of what it would take for a wolf to be rational, '(w)e cannot allow ourselves to suppose that God, say, might confer reason on wolves, but stop short of giving them the materials to step back and frame the question "Why should I do this?" '[4] If this is right, then appeal to SRR as governing a concept used by a rational thinker should be sufficient to 'individuate' first-person reference. But is the restriction on rational thinkers not just a trick to avoid counterexamples? This of course depends on our theoretical aspirations. If we are interested in explaining what has to be added to some creature's or god's 'use' of a referential device governed by SRR to turn this into a case of *thinking* first-personally, then (1') would of course be quite unsatisfactory. However, such reductive ambitions need not be part of the project of elucidating what it is to think first-personally. One's interest may be merely in what is distinctive of the first-person way of thinking of an object.

Up to a point, O'Brien agrees with this line of response. She notes that the two gods are 'in some peculiar way estranged from their own utterances' (p. 58) and suggests that this estrangement rules out 'the possibility that the gods are genuine speakers or thinkers' (ibid). Still, in her view, at least 'the case suggests that if SRR, as used by us, is sufficient to individuate first-person reference it is because of something we have that the gods lack' (p. 58). Yet, under her own diagnosis, it would be more nearly right to say that it is because of something we *are* that the gods are not, viz. genuine thinkers. Her insistence that we need to identify something we *have* that the gods lack—adding which would be sufficient to enable the gods to think of themselves first-personally—

3. See J. Campbell, *Past, Space, and Self* (MIT Press, 1994), p. 133.
4. J. McDowell, 'Two Sorts of Naturalism,' in R. Hursthouse, G. Lawrence, and W. Quinn (eds.), *Virtues and Reasons* (Clarendon Press, 1995), p. 153.

reflects a reductive ambition we have been given no reason to regard as compulsory.

3. Self-Consciousness

The thing we have and the gods lack, of course, is 'agent's awareness.' But what is 'agent's awareness?' I mentioned that O'Brien uses both transitive and intransitive constructions in talking about 'agent's awareness.' Such awareness involves, in some sense, experiencing an action or being aware *of* the action. At the same time, it is a matter of the action being a *conscious* action. I think the canonical account of agent's awareness is provided by the transitive construction, not just because it is more frequently used but also because it seems to be demanded by the task to which 'agent's awareness' is being put. The idea, recall, is that 'agent's awareness' is a (primitive) form of *self*-consciousness and is as such qualified to constitute an intelligible source of *first-person* knowledge.

O'Brien's proposal is this: 'we are agent-aware of our actions in virtue of carrying them out as a direct result of an active consideration of ways we might act' (p. 120). It is often said that we know what we are intentionally doing simply in virtue of the fact that our actions express our intentions. In Anscombe's phrase, agents' knowledge is 'knowledge in intention.' One might read O'Brien as endorsing a version of this idea. She quotes with approval Moran's Anscombean reflections on the connection between self-knowledge and deliberation. And the most natural way to interpret her frequent references to an agent's 'active consideration or evaluation' of 'possibilities for action, grasped as possibilities' is precisely to take this to involve a process of practical deliberation. O'Brien's theory, then, might be seen as a development of the traditional picture, though with two key differences.

One is that for O'Brien, what matters is a phenomenologically salient process of deliberation. If you spontaneously catch a ball I suddenly throw in your direction, on O'Brien's account, you will arguably lack 'agent's awareness' (there being no time to consider practical possibilities, grasped as possibilities), and accordingly, you will lack the usual source of first-person knowledge of what you are doing. Anscombe, in contrast, would insist that your action is intentional (hence, *intelligible* in terms of practical deliberation) and as such open to 'knowledge in intention,' even if the intention is acquired spontaneously, not preceded by deliberation.

The second, connected difference is that while on Anscombe's and Moran's view, there is a sense in which knowledge of one's own intentional actions is based on *practical* reasons rather than any sort of evidence—we acquire such knowledge by answering the practical question of what to do—O'Brien project is to identify a kind of evidence or theoretical reason that provides the source of such knowledge. She thinks the fact that an action is conscious predisposes it to 'stand as the reason for its own self-ascription' (p. 120). A noteworthy feature of this account is that knowledge of one's actions turns out to be merely an instance of the completely general point (quoted earlier) that

'there is a general entitlement immediately to self-ascribe those states and activities (. . .) which are conscious' (p. 119).

There is obviously much more to be discussed and probed in O'Brien's account, thus interpreted. It is time to acknowledge, though, that this reading cannot be quite right. Deliberation, as Anscombe and Moran understand it, involves answering questions such as 'what should I do?' or 'why should I do this?' If O'Brien's 'sense of control' is a matter of conscious deliberation, this would make it relatively easy to understand her view that it involves a form of self-consciousness. The self-consciousness in question would be the familiar variety associated with the use of the first person. But of course this is not what is required for O'Brien's theoretical purposes. O'Brien seems to think of agent's 'active consideration of possibilities for action, grasped as possibilities' as a more primitive phenomenon, involving a 'pre-first-personal' form of self-consciousness. She writes, for example, that 'agent's awareness' is a 'form of awareness which is such that a suitably cognitively equipped subject—a subject with grasp of the first person and the concept of an action—will immediately be able to self-ascribe the action they are conscious of in this way' (p. 188). Formulations like this suggest that 'agent's awareness' does not *require* possession of concepts, at least not of the concepts one might have thought to be relevant to grasping possibilities for action as such. Rather, the idea seems to be that one can give an account of what it means to grasp possibilities for action as such that invokes only non- or proto-conceptual materials. Let us grant this. Still, I want to question whether such an account could, in principle, help to achieve O'Brien's theoretical objectives.

Suppose first that the account agrees that even proto-deliberation is, in a sense, first-personal. For example, it might be suggested that we have a primitive ability to distinguish actions that are physically open to us from ones that are not, perhaps invoking some preconceptual body schema. So long as this ability involves proto-*first person* thought, however, the account would raise the question that is not supposed to arise, of what determines the reference of the 'foundational' first-person representation. In other words, the account would simply join the long list of crude and therefore unsuccessful attempts to implement the 'two-tier strategy.' As indicated earlier, O'Brien is quite clear that a properly refined version has to forgo any appeal to first-person reference in framing the second tier.

Suppose then that it is possible to explain what it means to 'grasp possibilities for action as possibilities' without appealing to anything resembling first-person reference. The trouble is that this would make it difficult to see in what sense a 'sense of control' could be called a primitive form of *self*-awareness. O'Brien is surprisingly casual about this. At one point, she writes that she does not mind whether we call the primitive form of awareness in question 'self-awareness' or something else (p. 188). The important thing, she suggests, is that it is 'self-indicating,' in the sense that it provides a source or basis for first-person thought. This is ingenious, but not satisfactory. If 'agent's awareness' is to be the basis of first-person thought, there has to be an *intelligible* link between it and use of the first person. What was supposed to render the link intelligible was precisely the idea that 'agent's awareness' constituted a primi-

tive form of *self*-awareness. We can then hardly appeal to the alleged fact that 'agent's awareness' provides a basis for first-person thought to vindicate its claim to be a form of 'self-awareness' or to be at least 'self-indicating.'

In summary, it seems to me that O'Brien's exceptionally sophisticated conception of what would be involved in making first-person thought philosophically intelligible in terms of a more primitive, foundational capacity for self-awareness helps to see more clearly why that project cannot succeed. At a rather high level of abstraction, the project may be compared to Hume's search for the origin of the idea of the self, and recalls Strawson's comment on that project: 'His attempt is to give an adequate explanation of the vulgar conception of the self as subject of experience; but the terms in which he conceives of such an explanation make it impossible for the attempt to succeed.'[5]

5. P.F. Strawson, *The Bounds of Sense* (Methuen, 1966), p. 170.