THE PROCEEDINGS OF THE TWENTY-FIRST WORLD CONGRESS OF PHILOSOPHY

VOLUME 6

Epistemology

EDITORS

Dermot Moran

University College, Dublin

Stephen Voss

Boğaziçi University, Istanbul

Philosophical Society of Turkey
Ankara 2007

The Sense of Agency and the Naturalization of the Mental

Costas Pagondiotis and **Spyros Petrounakos**

In this paper we examine whether the sense of agency represents an obstacle to the project of naturalizing the mental. On the basis of a thought experiment we suggest that the sense of agency is not an epiphenomenon. We also examine Frith's attempt to explain in functionalist terms the sense of agency through the comparator and metarepresentational mechanisms. Through a variety of arguments we try to show that explanation by recourse to these mechanisms is inadequate. We conclude by suggesting that one possible reason for the failure of the functionalist approaches is that they begin from the assumption that thought is a form of willed action.

In contemporary philosophy of mind there is a prevalence of theories that attempt to naturalize mental phenomena. Of these, the most basic are certain versions of functionalism and of the representational theory of mind. One standard objection to these projects is that by naturalizing mental phenomena they leave consciousness and qualia out of the picture. The usual reply to this objection is that qualia are epiphenomena and, as such, they just accompany mental phenomena without having any cognitive role. In this paper we want to raise a parallel objection to the naturalistic projects, to the effect that they leave self-consciousness out of the picture and that one cannot account for mental phenomena at a sub-personal level.

Self-consciousness is a cover term for many different characteristics of human mentality. We often use this term to refer to personal identity through time, to the unity of consciousness, to our ability to entertain thoughts as ours, and so on. In this paper we are going to focus exclusively on the sense of mineness or agency that characterizes conscious thoughts.

Let us start with a thought experiment. Suppose there are two identical twins A and B living in molecule-by-molecule identical environments. Let us also suppose that at time t both A and B are sitting in their identical rooms and entertaining the same thought, namely the intention to go to the kitchen to drink water. The thoughts are the same by both internalist and externalist criteria: both thoughts have the same narrow and wide contents. Now, if the sense of agency is really an epiphenomenon, then it follows that depriving subject B of that sense would make no difference to his behavior and his dispositions to behavior. That means that the behavior and the

dispositions toward behavior of subject B would continue to be identical with those of subject A, who has not been deprived of the sense of agency.

But would it be so? Not at all! The difference would be immense. If I were to lose that sense of agency from the thought to go to the kitchen to drink water, then I would experience a totally unfamiliar situation: I would find myself with a thought that I would not feel as mine. It would be as though someone else had put that thought "into" my mind, given that there are no such things as *orphan* thoughts. Moreover, this thought would be more like an order to go to the kitchen to drink water than an intention on my part to go to the kitchen to drink water. The presence of such an uncanny thought would, most probably, produce in me a feeling of terror and a desire to flee rather than to go to the kitchen to drink water. But even if I did the latter, this act would not be understood as deliberate but as coerced.

If a mental phenomenon is deprived of the sense of agency, then it is cut off from the mental stream of the subject, which, at a first approximation, can be thought of as a network of beliefs and desires. The particular mental phenomenon is not connected with the content of my beliefs and desires, nor can it be readily integrated with them. Thus, it occurs as an isolated thought that, in order to be rationalized, is attributed to somebody else and is experienced as an inserted thought or a heard voice. These are symptoms of schizophrenia. Therefore, returning to our thought experiment, the difference in the behavior and the dispositions toward behavior between A and B is as big as the difference between the behavior and the dispositions to behavior of a normal subject and a schizophrenic. Thus, the sense of agency is not an epiphenomenon. Rather, it plays a constitutive role in the organization of the mental life and behavior of the subject.

Some naturalists, however, attempt to give a functional explanation of the sense of agency rather than taking it to be an epiphenomenon. Christopher Frith (1992), for example, attempts to account for the sense of agency in terms of a cognitive mechanism, the comparator. This mechanism was used originally to explain the way we control our bodily movements and differentiate self-initiated movements from other movements. Frith maintains that thought can be understood as a form of action: thoughts, like actions, arise from prior intentions. As such, our ability to discriminate between a thought that is our own and one that is not can be similarly explained through comparators.

The comparator accounts for bodily action by comparing information produced by the intention to act with information received from proprioceptive and visual feedback concerning the movement that

has taken place. The comparator model is an attempt to explain our ability to discriminate between actions caused by our own goals (willed actions) and actions that are in response to external events (stimulus-driven actions) in terms of the presence or the absence of a motor instruction copy (of the intention to act) in the comparator system. Frith applies this idea to thought as follows: if a copy of an intention to think is not sent to the relevant comparator, the thought in question is not accompanied by a sense of agency and is experienced as 'alien'.

A first problem with the comparator model is that it seems to be insufficient as an explanation of how we come to experience an action as our own. This is because the comparator is essentially a monitoring device that is similar to those used in engineering for controlling a machine's operation. The point here is that it would be highly implausible to claim that such devices possess a sense of agency. It would be equally implausible to claim that every living organism that possesses such monitoring devices—like, for example, fruit flies—can be credited with a sense of agency.

An additional problem emerges if we question the need for such a mechanism for thought. This is certainly not a problem for the case of action. Unlike thought, our capacity to discriminate between willed action and stimulus-driven action has very practical consequences, e.g. in maintaining postural balance. John Campbell (1999, p. 616) attempts to meet this objection by suggesting that thought comparators serve the purpose of preserving the coherence of our thoughts, of keeping our 'thoughts on track, to check that the thoughts you actually execute form coherent trains of thought'.

The problem with this suggestion is that coherence is a semantic characteristic. Given this, it is not at all clear how the role of preserving coherence in thought can be allocated to a mechanism that supposedly has access to thought only at the syntactic level. This would only be possible for deductive arguments where the semantic property of logical validity can indeed be mirrored by the syntactic property of provability. Yet our thinking in everyday life does not, on the whole, have the structure of deductive arguments but is based mostly on inductive reasoning, on analogies and so on. This introduces a holism that prohibits the evaluation of non-deductive thinking on the basis of local features such as the presumed syntactic properties of thoughts.

A final problem is connected to the very idea of an intention to think. Though there is such a thing as an intention to think, e.g. in solving a mathematical problem, it is implausible to claim that, at the phenomenological level, all thoughts are preceded by intentions to

think these thoughts, in the same way that it might be said that an action is preceded by an intention to act. This problem also has a logical dimension: if we accept that a thought must always be the result of a previous intention, we end up with an infinite regress.

Campbell (1999) has suggested that the solution to this problem is to view the prior intentions as unconscious and thus as intentions that are not available at the phenomenological level. But even if we try to do justice to the phenomenology in this way, we will would still be faced with a more general problem, which is that it remains unclear how the comparator model can account for the misattribution of the sense of agency. According to this model, if there is no matching between intention and thought, thoughts should be experienced as merely orphan, as lacking any agency whatsoever. But inserted thoughts are not experienced merely as orphan but as alien, as due to another *subject*.

A functionalist might respond by claiming that this attribution of agency to another subject is due to another mechanism, a mechanism of metarepresentation. Such a metarepresentational mechanism has also been suggested by Frith. The main idea is that the misattribution of the sense of agency can be explained in terms of an abnormality in the capacity to metarepresent, that is, in the capacity to produce second-order thoughts that endow first-order occurrent thoughts with a sense of agency. But this mechanism would lead to an infinite regress: recognizing a first-order thought as our own in virtue of a second-order thought entails that such a second-order thought would itself require a third-order thought in order to be recognized as ours, and so on.

Here one could attempt to prevent the occurrence of an infinite regress by saying that it is only the recognition of first-order thoughts as our own that we seek to explain and that there is no need for a corresponding explanation of the second-order thought. But this is an implausible response for two reasons. Firstly because the resulting picture would contain thoughts—albeit of a second order—that would still be alien and secondly because it would entail that these alien thoughts, qua alien, could nevertheless be responsible for the sense that the first-order thoughts are ours.

Thus far we have seen that subpersonal mechanisms do not suffice to explain the sense of agency. We would like to suggest that one reason for this failure is the assumption that thought can be explained on the model of willed action. This assumption leads to circularity. The reason is that willed actions are distinguished from stimulus-driven actions by recourse to an intention—that is, to a kind of thought—that the former involve. Thus, ordinary thoughts

cannot be distinguished from inserted thoughts by recourse to willed actions (which are distinguished from stimulus-driven actions by recourse to a thought) without circularity.

There is also a second reason that undermines the analogy between willed action and, as it were, 'willed' thought. Whereas in the case of willed action, having an intention to act is having in view what the action will be, no such thing happens with willed thought: even the most typical case of an intention to think—e.g., when we are trying to solve a problem—does not involve having in view the solution that we will eventually think of.

A more general problem here is that this assumption has Cartesian roots and as such leads to a mentalistic picture of the entire issue. We believe that within such a picture the chances of arriving at an adequate account of the sense of agency are slim. We would like to end with the suggestion that we might be better placed to investigate the issues above if we were to view thinking as a kind of skillful activity. The main reason here is that such a description of thinking does not involve mentalistic terms, which have their origin in a Cartesian perspective.[1]

Costas Pagondiotis, Department of Philosophy, University of Patras, 26 500 Rio - Patras, Greece. E-mail: cpagond@upatras.gr
Spyros Petrounakos, 5 Agiou Isidorou Street, 11471 Athens, Greece. E-mail: spypet2@yahoo.co.uk

REFERENCES

Campbell, John (1999). "Immunity to Error through Misidentification and the Meaning of a Referring Term." *Philosophical Topics* 26.

Frith, Christopher (1992). *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pagondiotis, Costas and Spyros Petrounakos (2002). "First Person and Thought Insertion" (in Greek). *Deukalion* 20/2.

NOTE

1 We elaborate on this suggestion in Pagondiotis and Petrounakos (2002).