

When self-consciousness breaks

G. Lynn Stephens & G. Graham

Cambridge, Mas. : MIT Press, 2000

This fascinating book deals with the experience of externality, i.e. an experience characteristic of verbal hallucination and of thought insertion, in which a subject attributes to an external cause an event that in fact occurred in her own mind. In verbal hallucination, a subject hears voices: it is part of her experience that these are persons speaking to her (usually in a critical manner). The oddity of such an experience is that the hallucinated subject does not have any visual experience congruent with what she hears. In thought insertion, the oddity goes still deeper: here, a subject disavows being the original thinker of one particular thought. As philosophers of mind have often remarked, it is difficult to make sense of the report on such an experience, for “thinking someone else’s thought” seems to describe an impossible situation. The patient insists however that some thinking episode feels being externally driven, such that the corresponding thought looks literally implanted in her mind.

Whereas thought insertion is a symptom of a schizophrenic disorder listed in DSM IV, verbal-auditory hallucinations typically affect schizophrenic patients as well as ordinary people – a survey reports that 70% of a sample of college students experienced them at least once. The analysis of this phenomenon is accordingly not conducted here as part of an attempt to understand schizophrenia; it rather aims at unravelling the psychological structure of human consciousness. Even though, as the authors finally admit, much remains to be learnt on self-consciousness, the book offers two important contributions. One is a valuable discussion of the various ways of addressing the paradoxical experience of externality. The other is an emphasis on a distinction between the experience of subjectivity and the experience of agency. This review will try to show that this distinction is indeed a crucial feature in any solution to the question of externality, but that it is associated with a view of thinking as acting that is questionable.

The authors themselves see their book as polyphonic, involving with several voices (see p. 133), so that we can only recover the argument of the book by following their variations on a central theme. What constitutes the book's value, and main interest, is that it offers a critical survey and a lively discussion of much of the contemporary literature on the subject.

The authors first present a theory for verbal-auditory-hallucinations -- developed in various versions, by Thomas Szasz, Louis Gould, Marcel Kinsbourne, -- which they call SPM, for *the self-produced but misattributed conception*. This account is based on two claims. A patient who hears voices generates the messages expressed by the voices and misattributes them to another person. One way of framing this account hypothesizes that a patient hears one of her own subvocalized inner speech sequences as someone else's utterance. Obstructing subvocalization in various ways (like holding one's mouth wide open, or humming a note), seems in most cases to block verbal hallucinations. Now what does one mean by "hearing one's voice as someone else's"? The so-called "auditory-hallucination model" (AHM) is a claim common to SPM theories, according to which a crucial factor in misattributing the voice's message is the similarity of the experience of hearing this particular voice and the typical experience of hearing another person speak. The hallucinator *confuses* what he just imagined with what he might, in other circumstances, have perceived, because of a close similarity in the corresponding experiences. At this point, a circle needs to be avoided, explaining similarity between experiences by a tendency to mistake one for the other, and explaining the tendency to mistake one for the other through the similarity of the experiences. Perhaps, as the authors also suggest (p. 47), the similarity in the experiences is not of a *phenomenal* nature; it is rather restricted to their properties of having one definite content (the fact that it is verbal-auditory, independently of the particular qualitative feel of such a content) and of not being under voluntary control. But this needs to be substantiated, and this is where Ralph Hoffman's 1986 paper in *BBS* offers promising suggestions.

In agreement with an influential research by Maher (1974), Hoffman argues that misattribution is not a result of a misguided reasoning or of a disturbed thinking process, but rather of altered perceptual data. Hoffman's specific contribution includes two claims: a) The voices are verbal images that, in contrast with ordinary experiences of inner speech, are perceived as unintended and b) unintended speech is automatically taken to be alien. Claim *a* is justified by the view that schizophrenic patients suffer breakdowns in the planning processes of their outer and inner speech; (this hypothesis has been well documented in various experimental data by Frith, Jeannerod and their respective co-workers¹). The impression of unintendedness is plausibly a result of this defective prediction of auditory feedback. Claim *b* is obviously much more difficult to establish. According to Hoffman, a patient experiences her inner speech as "strongly unintended", i.e. not only in the sense that

¹ See Frith (1992), Frith et al. (2000), Daprati et al. (1997), Jeannerod . (1999).

she does not have a conscious access to the corresponding plan – this is the “weak sense”-, but also in the stronger sense that its occurrence actually conflicts with her occurrent conscious goals and values. In such a case, a patient might be unable to handle the proper reality test that, in the weak case, normally allows inhibiting the non-self inference.

The authors are not fully satisfied with Hoffman’s theory, because it leaves some crucial points unexplained. In particular, why should unintended inner speech be automatically interpreted as externally produced ? Furthermore, is it not difficult to believe that unintended inner speech should be related to a *discourse planning failure*, as the voices are generally well structured and have clear communicative goals ? Why are the voices always addressing the patient herself ? And how might Hoffman’s account deal with the numerous cases in which voices are experienced as both *internal* (rather than *audition-like*) and *alien* ?

The authors aim at building an alternative account, where the verbal-auditory hallucinator does not simply *mistake* inner speech for auditory perception, and where voices do not *need* to be externally perceived. In this second approach, voices and thought insertion are taken to be closely related phenomena that call for a common account. Thus the discussion moves to another clinical aspect of the schizophrenic syndrom : how is thought insertion to be explained ?

Philosophers diverge on the issue of whether access to one’s mental happenings necessarily implies their attribution to oneself. The problem, sometimes named the “transition problem” (Peacocke, 1999), involves understanding the kind of link that allows a thinker to derive, from having a certain perception P or a thought T, the belief that [I perceive P] or [I think T]. Locke, Chisholm and Shoemaker defend the view that the transition link is necessary, with arguments that range from metaphysical to functional/ epistemological claims. Other philosophers hold that the transition is far from necessary. Hume famously rejected the very possibility of a natural transition towards a self in whom ideas are supposed to occur. In a similar vein, Armstrong (1968) holds that the self referred to in "I think T" is a theoretical construct that goes beyond the data given in awareness. Using Armstrong’s view, one might hold that a thought can be self- or other-attributed only if two conditions are satisfied : 1) a thinker has acquired verbally the notion of a self (along with a theory of mind), and 2) she has introspective access to occurrent mental happenings. When the latter are coherent with her long-standing beliefs, desires and intentions, she is able to infer that such thoughts are her own, i.e., are produced in her mind; when they are not so coherent, in certain further conditions to be spelled out, the inference might be that “they are not states of

the same substance” (see Campbell, 1999, for a similar view applied to perturbed identity in schizophrenic patients).

Such an account would miss a crucial fact, according to the authors. For a deluded patient may experience a particular thought as alien while also insisting that this thought occurs in *her own mind*. A patient who experiences thought insertion might thus seem to have retained a sense of her ego-boundary. One might accommodate Armstrong’s theory however by contrasting the conception of her mind acquired on the basis of inferences from *earlier* normal episodes of thought processes with her *present* impression of extraneity, linked to the incongruence of an *occurrent* thought or train of thoughts. A patient could thus well retain her former sense of a proprietary mind (based on the previous coherence of her thoughts), while also insisting that thoughts that present themselves in it now are not her own (for the new thoughts fail to be coherent with her acquired dispositions). In fact, this kind of solution of the experience of alienation is close to the one that the authors will eventually offer. We will come back to it below.

Let us grant Stephens and Graham, at this point, that explaining the subjectivity of our thoughts does not exhaust the question of ownership recognition, for we need to explain what the additional conditions are in which a subject will attribute a thought to another mind rather than simply disown it. Armstrong’s theory needs to be supplemented by an account on the kind of monitoring needed for self- or other-attribution of thoughts. Frith’s account offers this additional piece of theorizing : voices and delusions of thought insertion reflect an impairment in action monitoring. A patient hears voices because she does not recognize her intended actions as her own. This account of thought-insertion is based on the view *that thinking is acting*: a patient with this delusion does not have any "sense of effort" associated with having a particular thought. This account differs from Hoffman’s by claiming that the patient did intend his words or thoughts (for Hoffman, remember, the words and thoughts were unintended as a result of a defective discourse planning device.). What Frith’s account still does not explain however, as the authors show, is why the failure in intention monitoring results in a sense of alienation rather than simply in a feeling of unintendence. Neither does it explain why a subject so perturbed may end up having a delusion of control (where someone influences his actions or thoughts) rather than a delusion of thought insertion.

Harry Frankfurt (1988) is then called as a third voice to complete Hoffman’s and Frith’s intuitions with a conceptually adequate notion of externalization: my arm can go up without my wanting it (for example if a neurosurgeon activates the relevant neurons); just the same,

some thoughts can occur in my mind without my having formed any intention to think them. These are thoughts that I find occurring in me, rather than thoughts in whose occurrence “I actively participate”. A patient’s delusional experience can be understood on the basis of this new distinction. A patient may well claim that a thought occurred in her mind while also insisting that it was not *her* thought. These two claims refer to the two previously established dimensions in self-awareness, the *sense of subjectivity* (by which a thinker recognizes having the experience of thinking a particular thought), and the *sense of agency* (by which a thinker recognizes – or fails to recognize – that she is actively entertaining that thought). Now our initial problem surfaces again : how can a patient's sense of agency be disturbed to the point of attributing to someone else her own thoughts and intentions ?

Two hypotheses seem available. The first is that lack of voluntary control forces the experience of alienation. But this assumption seems immediately defeated by evidence: people with obsessive compulsive thoughts – who have no voluntary control on their thoughts - do not tend to attribute their thoughts to other thinkers. A second hypothesis may be derived from Dennett’s view on the self . Let us grant that the self is an entity in a narrative, produced as a result of a culturally-educated attempt to organize in intentional terms an agent’s past experience and decisions. In that vein, one might claim that a patient denies having one particular thought because it does not fit her conception of herself. This solution is very close to the elaboration on Armstrong's view articulated above. One significant difference is that the contrast between subjectivity and agency/passivity is superimposed on Armstrong’s contrast between a previously acquired conception of oneself and new congruent/non-congruent mental happenings. What is called subjective is what fits, what is called objective or external is what does not. Still, why should a patient attribute the non-congruent thought *to someone else* ? The long-awaited explanation of alienation that is offered in the last pages of the book seems disconcertingly plain. Here is the main passage :

“As far as phenomenology itself is concerned, our hypothesis is that the apparent intelligence of the thoughts provides the experiential or epistemic basis for attributing them to another agent. Mary [a subject having an experience of alienation] experiences her thoughts as “personal” (intelligently composed by someone), but not as expressive of her own person”. (174).

This explanation, the authors say, is close to Hoffman’s by assuming that the experience of alienation is epistemically warranted given the evidence available (i.e. as of an intentional action foreign to the subject’s own goals); it is also close to Frith’s by offering a unified explanation for hearing voices and thought insertion. Finally it is close to Frankfurt’s by

assuming that people are agents qua thinkers and not only qua physical forces. It further has the merit of distinguishing conceptually the case of schizophrenic deluded patients with thought insertion, who deny having formed a thought, from obsessive thinkers, who take the blame for thinking their compulsive thoughts. It also distinguishes nicely thought-insertion from thought control. In the latter case, a subject feels manipulated in her agency; she retains a capacity of forming intentions, but takes her intentions to have been fiddled with; in the former case, in contrast, she takes her agency to have been completely taken over by someone else; it is part of her experience that no intention of hers has been formed or distorted.

The authors' explicit aim is to offer a conceptually adequate description of the types of experience involved in alienation; they emphasize accordingly that the contrast between a sense of subjectivity and a sense of agency plays a central role in the phenomenological aspects of pathological states. Even in those cases where patients attribute their own acts to some external force, they experience agency-deprivation in a first-person, "subjective", way. This difference in the dimensions of self-awareness can also be found in other psychopathological conditions not considered in the present book. Depersonalized patients also report having a subjective feeling of depersonalization, and patients with Cotard syndrome that they do not exist anymore. The fact that a subject could retain a first-person experience of episodes that seem associated with an impression of passivity or of alien influence is certainly a major indication in favour of a polarity in conscious awareness between simple conscious registering and self-attributing an active/passive role in acting and thinking. The conceptual contribution offered however is not intended to provide a causal explanation of the various symptoms discussed. The reader may at this point have an impression that the rules have changed, for Hoffman and Frith were criticized for offering no *causal* explanation of the difference between, say, lack of control and sense of alienation, and not for lacking sensitivity to the conceptual opposition between the two varieties of symptoms.

Let us come to the proposal itself, namely that thought insertion is a case of failed agency, experienced by the agent at a personal level as an intelligible thought to which she cannot identify ("not expressive of her own person"). This explanation is valuable as it stands, in particular because it explains the patient's perspective on agency in epistemologically clear terms (the subject is not irrational, her ways of deriving intentionality and goal-directedness are similar to everyone else's.). Two difficulties however seem worth articulating and exploring further. The first is linked to the assumed contrast in the quotation above, between an "intelligently composed thought" and a "thought expressive of her own person". In other words, the subject is able to recognize a complex thought, which triggers a mechanism of

interpretation (who is the thinker ?); but she is unable to match what she hears or grasps with a thought she produced. One problem for the explanatory value of this contrast (intelligible but incongruent) is that the very capacity to appreciate composition in thought seems to be quite similarly used in understanding and in producing a thought. The fact that some structured utterance is heard or mentally grasped "as an intelligible thought" does not bias the attribution one way or another. One would rather suppose, on a conceptual basis, that the fact is neutral as to who said what. The incongruent content, taken at the phenomenological level, can be dealt with in other ways than in projecting the corresponding intention into another agent; for example, as an isolated memory popping up, or as a piece of unexplained compulsive thought; incongruence does not seem to constrain interpretation towards alien intrusion.

An alternative claim would be that the feeling of agency is not causally closed on phenomenology. It is not inconceivable, given the growing neurophysiological evidence, that the subject described is biased by a specific tendency to over-attribute properties to others - an impairment that might cumulate its effect with a general difficulty in appreciating the difference between active, voluntary thinking/acting and automatic association of thought/movement sequences. Such a difficulty might be generated by a perturbed subpersonal mechanism balancing self-other attributions (through a regulation of cortico-cortical projections, effected by structures such as the prefrontal area and the right inferior parietal lobule). If such a bias exists, then it is hopeless to try and interpret the delusion on the basis of some *other* features *apparent in the phenomenology*.

A second problem lies in a tension between two lines of explanation, with a resulting ambiguity concerning the status of our thoughts. One is the view initiated by Hughlings Jackson and Irwin Feinberg, developed by Christopher Frith, and defended more recently by John Campbell (1999) : thinking is a type of action, involving motor activity and monitoring of intentions; a mechanism of efference copy such as the one that subserves action awareness has to be present both to plan one's trains of thought and to retrospectively attribute them to oneself. The other is the view, inspired by Frankfurt, according to which thinking is not immediately an active process; it can however be appropriated through the exercise of second-order thoughts. A subject may thus act or think in a way that does not reflect what she wants to do: she may be "irresistibly inclined" to do or think something. Let us note that thinking, in that view, is not equated with action; both thinking and acting may be passive or active. Frankfurt does not think that a subject *normally* has control on all his thoughts or intentions to act; he takes it that a subject will at best only identify with a subset of these. If and only if he

is able to form second-order thoughts, desires and intentions, he will be in a position i) to chose those of his first-order thoughts, desires and intentions to which he wants to identify; ii) to appreciate their coherence with his second-order preferences.

What makes the parallel between Feinberg/Frith and Frankfurt uneasy is an irreducible difference in levels of explanation ; the first account explores subpersonal mechanisms regulating the way in which a subject may become aware of what she does. The second analyses the reasons that a person may have to feel free in a deterministic world. The issue is to show that there is more to autonomy than just having thoughts or acting out of desires; it is identifying oneself "actively" with them. It makes no sense, in the latter perspective, to raise the question of the experience of autonomous agency in a subject who would not grasp that a thought is occurring in her mind, or that she performed a certain action. The subject must at least recognize that he involuntarily thought that thought or did that action. Therefore Frankfurt's kind of approach does not have any bearing on the deluded patient's case.

Furthermore, how can the view that a thought can only become "active" when a second-order thought is used to evaluate or guide its first-order target, be reconciled with a theory in which *every* thinking episode registers as an action via an efference copy ? We have here a bootstrapping problem : only when reaching a second level of awareness can a sense of autonomous agency be gained. But some kind of conscious active agency must already be in place for the higher kind of agency awareness to get started.

Maybe a central difficulty consists in assuming with Frith that thinking is in all cases a variety of acting. One might rather claim that a thinking episode qualifies as a mental act only if it involves a controlled process guiding the thinker to a target mental property, a definition that is close to Frankfurt's own view. Most of our ordinary thoughts do not qualify as actions in that sense, just as many physical movements are done with no intention in mind. If we thus have two distinct concepts of active thought in Frith and in Frankfurt, the latter's account cannot "dovetail" on the former's distinction (see p. 152).

The book under review offers much more than a particular analysis of the experience of alienation; it is a lively and documented philosophical introduction to difficult questions on the pathologies of consciousness and agency. It is certainly one of the first endeavours of this kind within analytic philosophy and should help many students and researchers to discover the new field of cognitive psychopathology as well as the important philosophical issues that arise in it.

Joëlle Proust

References

- Armstrong, D., (1968), *A Materialist Theory of the Mind*, London, Routledge and Kegan Paul.
- Campbell, J., (1999) Schizophrenia, the space of reasons, and thinking as a motor process, *The Monist*, vol. 82, 4, 609-625.
- Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J. & Jeannerod, M., (1997), Looking for the agent, an investigation into self-consciousness and consciousness of the action in schizophrenic patients, *Cognition*. Vol. 65, pp. 71- 86.
- Feinberg, I., (1978), Efference copy and corollary discharge : implications for thinking and its disorders, *Schizophrenia Bulletin*, 4, 636-640.
- Fourneret, P. & Jeannerod, M., (1998), Limited Conscious monitoring of motor performance in normal subjects, *Neuropsychologia*, 36, 11, 1133-1140.
- Frankfurt, H.,(1988) *The importance of what we care about*, Cambridge, Cambridge University Press.
- Frith, C.D., Blakemore, S.-J., & Wolpert, D.M., (2000), Explaining the symptoms of schizophrenia : Abnormalities in the awareness of action, *Brain Research Reviews*, 31, 357-363.
- Frith C.D., (1992), *The cognitive Neuropsychology of Schizophrenia*, Hillsdale, Lawrence Erlbaum Associates.
- Hoffman, R., (1986), Verbal hallucinations and language production processes in schizophrenia, *Behavioral and Brain Sciences*, 9: 503-517.
- Jeannerod, M., (1999), To act or not to act, perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology*.52A :1-29.
- Maher, B., (1974), Delusional thinking and perceptual disorder, *Journal of Individual Psychology*, 30, 98-113.
- Peacocke, C., (1999), *Being Known*, Oxford, Clarendon Press.