THOUGHT INSERTION, SELF-AWARENESS, AND RATIONALITY

JOHANNES ROESSLER

Thought insertion raises two kinds of philosophical issues: general issues concerning the interpretation of what some diagnostic systems in psychiatry call bizarre delusions, and specific issues concerning the nature of self-awareness and its pathologies. The basic philosophical puzzle generated by the delusion, I suggest, may be put in the form of an inconsistent triad that brings together both kinds of concerns.

Transparency: To be introspectively aware of a current episode of thinking that p is to be aware of oneself thinking that p.

Alienation: Patients with the delusion of thought insertion believe that someone else is the thinker of an episode of thinking of which they are introspectively aware.

Intelligibility: Reports of thought insertion express rationally intelligible, coherent beliefs.

Let me start by going over the apparent inconsistency between the three claims. Consider an example of thought insertion: “Thoughts are put into my mind like ‘Kill God.’ It is just like my mind working, but it isn’t. They come from this chap, Chris. They are his thoughts.” Transparency—or its natural extension to the case of thoughts in the imperative mood—suggests that insofar as the patient is introspectively aware of an episode of thinking “Kill God” he is aware of himself thinking “Kill God.” His introspective awareness cannot settle the question of what is being thought without simultaneously settling the question of who is thinking it, viz. he himself. Then if the patient nevertheless believes that someone else is the thinker of the episode in question, as Alienation maintains, this would not just be an eccentric or mildly irrational belief, on a par with ideas such as telepathy. Rather the belief would be manifestly inconsistent with the content of the patient’s introspective knowledge. This interpretation would scupper any prospects for attempting to make rational sense of the delusion, as demanded by Intelligibility.

The standard view in the recent philosophical literature on thought insertion is that reflection on the delusion should lead us to discard, or at least modify, Transparency. This view goes back to two pioneering papers published in the 1990s, by Stephens and Graham...
THOUGHT INSERTION, SELF-AWARENESS, AND RATIONALITY 659

(1994) and by John Campbell (1999). They propose different versions of what I will call the two-concept view. Commonsense psychology, they claim, recognizes two distinct concepts of ownership of a thought, “introspective” and “agentive” ownership (my terminology, not theirs). Given that these are distinct concepts, one may coherently acknowledge one’s “introspective ownership” of an episode of thinking while simultaneously questioning or denying that one is the “agent” of the episode. This, according to the standard view, is what patients are doing. As far as this core element of the delusion is concerned, the delusion involves a perfectly coherent belief. Moreover, two-concept theorists standardly argue that it is possible to make sense of a patient’s acquisition of the belief by adverting to their highly unusual inner experience, perhaps involving an impaired sense of agency. This does not mean that proponents of the standard view will accept Intelligibility without qualification. They are not committed to the simple scheme “abnormal experience/rational belief.” They may concede, for example, that patients’ tendency to attribute the thinking in question to other individuals and to invoke abstruse mechanisms of “transmission,” or even just their apparent inability to appreciate the implausibility, all things considered, of their denial of “agentive ownership,” reflects a disorder of rationality, in addition to disturbed inner experience. However, these are relatively minor qualifications. The standard view does endorse Intelligibility at least on this minimal reading: the core of the delusion—the denial of ownership of an episode of thinking—is a coherent belief that is open to reason-giving explanation; patients’ abnormal inner experience provides them at least with a prima facie reason for denying “agentive ownership.”

I started by distinguishing two kinds of issues raised by thought insertion, to do with the interpretation of “bizarre” delusions in general, and with the nature of self-awareness in particular. I want to suggest that on both counts the standard view faces serious problems. In the first half of the chapter I focus on self-awareness. The standard view insists that self-awareness exhibits a certain explanatory structure, which in turn yields a substantive explanation of the kind of alienation voiced in reports of thought insertion. I want to argue that recent attempts to identify some such structure have been unsuccessful. I shall also suggest that there are good reasons to think the search for it is futile, and that Transparency resists revision (second and third sections, “Inner speech” and “Making sense”). In the fourth section (“Empirical versus autistic-schizophrenic delusions”) I present independent reasons for thinking that it is Intelligibility that should be modified. On the “historical” analysis of thought insertion to be explored in the last section (“Thought insertion and its history”), interpreting the delusion requires paying close attention to its prehistory in the prodromal phase of schizophrenia. On this account, the truth in Intelligibility is this: Reports of thought insertion express attitudes that originate in rationally intelligible, coherent beliefs.

**TRANSPARENCY**

I am going to use the notion of introspective awareness in a fairly non-committal way. To speak of introspective awareness of episodes of thinking, as I understand the notion, does not commit one to the view that introspection is a special source of knowledge, let alone one that involves some quasi-perceptual awareness of inner objects. Introspective awareness
of an event may just be a matter of having a distinctive kind of propositional knowledge of the occurrence of the event—knowledge, at a first pass, that is non-inferential and non-observational. We can take transparency in an equally low key. The basic idea is that when you enjoy introspective awareness of an episode of thinking, the question of who is doing the thinking is "transparent to" the question of what is being thought. That is, insofar as your introspective awareness provides an answer to the latter question, the first question is settled—you are committed to a self-ascription of the episode.

The idea is familiar from discussions of “immunity to error through misidentification.” For Transparency suggests that a certain kind of question, and a certain kind of error, would not be rationally intelligible. Consider Wittgenstein's well-known remark in the Blue Book (I've adapted the example):

There is no question of recognizing a person when I say "I think it will rain." To ask "Are you sure that it's you who are thinking it will rain?' would be nonsensical. Now, when in this case no error is possible, it is because the move which we might be inclined to think of as an error, a "bad move," is no move of the game at all.1

If Wittgenstein is right, then thought insertion, construed on the lines of Alienation, involves a belief that is not an error or a "bad move" but no move at all. This is what the two-concept view denies. Two-concept theorists agree that it would be “no move at all” to question one's introspective ownership of a current episode of thinking, perhaps by asking “Is this episode occurring in my mind?” After all, they hold that introspective awareness of an episode is a necessary and sufficient condition for ownership in the “introspective” sense. But they insist that "I think" is naturally used to talk about ownership in the “agentive” sense. And when it comes to “agentive” ownership, they claim, we should reject Transparency.

Here is how Stephens and Graham put the idea:

There may well be a sense of the verb to think in which the statement “I am the subject in whom the thought occurs” necessarily entails “I think the thought.” We contend, however, that the verb also has a sense in which "I think a thought" indicates not that I am the subject in whom the thought occurs but that I am the agent of the thought. It is this sense of to think that is relevant to understanding thought insertions. And when to think is used in this agentive sense the above entailment fails to hold. (Stephens and Graham 1994, p. 6)

And here is a series of quotations from John Campbell:

At the very least, these reports by patients show that there is some structure in our ordinary notion of the ownership of a thought which we might not otherwise have suspected.

[T]he schizophrenic has introspective knowledge of a thought of which he does not recognize himself to be the agent.

The content of the schizophrenic's illusion is that he has first-person knowledge of token thoughts which were formed by someone else. (Campbell 1999, pp. 610, 619, 620)

What does it mean to be the “agent of a thought”? A phrase commonly used in the literature as a more familiar-sounding variant is “author of a thought.” We know what is meant

---

1 See Wittengstein (1958, p. 67; Wittgenstein's example is “I have toothache.”)
by authorship of books or peace plans, and it seems unproblematic to speak of someone's authorship of a thought in just this sense. Note, though, that in the familiar sense, the term applies to types, not tokens. It is books that have authors, not individual copies of books. In the context of thought insertion, therefore, the term “authorship” is not particularly useful: it’s obscure what, if anything, the phrase “author of a token thought” might mean. This raises a further question. Are there such things as token thoughts? If seven members of the audience think the lecture is soporific, how many thoughts are we talking about? The natural answer, I suppose, is “one thought.” Could we say, in addition, that seven thoughts “occurred”? No doubt we could—Stephens and Graham, for example, use this construction in the earlier quotation. But that usage is arguably uncommon in ordinary parlance. Moreover, it breeds confusion. What is the relation between a thought someone is thinking and a thought that is “occurring”? The natural answer is that the former is the content of an episode of thinking, whereas the latter is the episode of thinking itself. Stephens and Graham tend to conflate the two notions: their wording suggests that when someone thinks a thought, that very thought is something that “occurs.” Partly to avoid the risk of equivocation, I prefer the cumbersome phrase “episode of thinking” to “token thought.”

How, then, should we understand the notion of “agentive” ownership? On what I’ll call the explanatory model, A is the subject of an episode of thinking, in the “agentive” sense, if the episode is intelligible in terms of A’s propositional attitudes. Stephens and Graham offer a version of this model, as does Campbell. More recent work in the two-concept tradition tends to exploit the idea that thinking involves acts of inner speech: “agentive” ownership of an episode of thinking, on this model, is a matter of being the agent of an act of inner speech involved in, or identical with, the episode (Byrne 2011; Jones and Fernyhough 2007; Langland-Hassan 2008). The question whether some version of either—or perhaps both—of these models can be sustained is a major issue in the philosophy of mind. It amounts to nothing less than the question of whether to think is to act, a question that can only be properly addressed in the context of a discussion of an even larger subject, viz., the question of what actions are. Fortunately, for my purposes here it will not be necessary to address these big issues head-on. For the question I want to press concerns not so much the notion of “agentive” ownership as such but rather its relation to the putatively distinct and independent notion of “introspective” ownership. What I want to challenge is the assumption that it’s possible to be introspectively aware of the content of an episode of thinking without being aware of being the agent of the episode. If the assumption is invalid, Transparency holds for both “introspective” and “agentive” ownership. Indeed no useful purpose will be served by distinguishing the two notions. I begin with the inner speech model. In the “Making sense” section I review the explanatory model.

2 Should this lead us to challenge Alienation? That is, should we question the near-universal assumption that patients use the word “thought” to refer to an episode of thinking? I think the matter is certainly less clear-cut than advocates of the standard view tend to assume. I return to the issue in the last section.

3 Another reason is the following. Is the thought shared by the seven thinkers adequately understood as a type of which the seven episodes of thinking are tokens? I doubt that it is. The type in question would have to be a type of event, but that is arguably not how we think of thoughts when we think of them as things we think or have. For illuminating critical discussion of the application of the type-token distinction to the (in some ways analogous) case of beliefs: see Steward (1997, chapter 4).
Inner Speech

Inner speech involves both agency and sensory phenomenology. This dual nature might inspire the following proposal. If you were to enjoy an experience with the distinctive phenomenology of inner speech without any awareness of performing an act of inner speech, you would be introspectively aware of an episode of thinking without being aware that it is you who is doing the thinking. You might then consistently, and intelligibly, affirm, say, that the thought “Kill God” is being entertained in your mind while denying that it is you who is thinking “Kill God.”

I think the most promising way to develop this proposal would be on the following lines. To say something in “inner speech” is to imagine saying something. This is not the same as imagining hearing anyone say something. Yet it involves a distinctive, and distinctively elusive, kind of auditory phenomenology. For as part of the imaginative project one imagines the auditory experience attendant on the imagined utterance. On the basis of that phenomenology, it might be said, you may recognize that an episode of thinking is occurring, even as you lack any awareness of the activity of imagining saying something. In an illuminating paper that provides the most detailed formulation of the “inner speech model” I’m aware of, Peter Langland-Hassan puts the idea as follows: “Because the sensory character of inner speech is not nearly as rich as that of actual speech perception, an act of inner speech is usually phenomenologically distinguishable as inner speech partly due to this paucity of sensory character” (2008, p. 393).

One worry about this proposal is that the various kinds of sensory phenomenology relevant here may not actually exhibit any such “Humean” gradation. Note that to identify an experience as an act of inner speech you would need to recognize it not only as imagining rather than hearing (or hallucinating) speech, but also as one of imagining a performance rather than an observation of a speech act. Especially in the case of the latter pair, it’s not clear that its members are clearly discriminable on the basis of their differential “paucity of sensory character.”

More importantly, I think it’s evident that we don’t actually rely on the character of the sensory phenomenology involved in inner speech to recognize that the experience is an episode of thinking. Rather, we know that an episode of thinking is occurring in virtue of the way we know what is being thought. Consider first the case of hearing someone else say “Kill God.” In that case, to know what thought is being expressed you may, and normally need to, listen to what he is saying. At the very least, auditory attention is a possible source of knowledge of the content and type of the utterance. Next suppose you imagine hearing someone

---

4 I think a third distinction is relevant too, between acts of inner speech that constitute episodes of thinking and acts of inner speech that do not. If I ask you to imagine saying “2 + 2 = 5,” your act of inner speech would not be naturally characterized as a case of thinking that 2 + 2 = 5. The point here is not that thinking requires belief. As Hampshire reminds us, “there are countless thoughts that occur to me, and that pass through, or that linger, in my mind, and of these only a small minority constitute beliefs” (Hampshire 1965, p. 97–98). The distinction we need is a more subtle one. One suggestion might be that episodes of thinking aim to express views that are least serious contenders for beliefs. If Hampshire is right, Byrne’s “epistemic rule” for self-ascribing thoughts—“If the inner voice says that p and p, believe that you are thinking that p” (2011, p. 121)—is too restrictive.
else say “Kill God.” In that case, your knowledge of the content and type of the imagined utterance will have no perceptual basis. You will normally have non-observational knowledge of what it is you are imagining. However, it is part of the imaginative project that the imagined experience is one that, in the imagined situation, would provide you with a possible source of knowledge of the content and type of the imagined speech act. Moreover, in the case of relatively spontaneous, “unbidden” imaginings it may be for the subject as if she came to know which speech act she is imagining by listening. Compare and contrast the case of imagining saying “Kill God.” In this case, you have non-observational knowledge of the type and content of the imagined speech act; and you are not imagining an experience that could provide you with a source of such knowledge—whether in the imagined or indeed in the real situation. The project of finding out what one is thinking by “listening” to one’s own inner speech is arguably just as fraught with difficulty as the project of finding out what one is saying by listening to one’s own current outer speech (see O’Shaughnessy 1963).

I think few would be prepared to deny the existence of some such epistemic “asymmetry” or deny that we have non-observational knowledge of the content of any episode of thinking of which we are introspectively aware. It seems to me, though, that once this point is granted, the idea of introspective awareness of a thought without awareness of being its thinker is beginning to look like a chimera. What is involved in having non-observational knowledge of what one is saying in inner speech? The natural answer is that such knowledge reflects the fact that acts of inner speech are intentional actions and as such performed “knowingly.” Following Anscombe we might call this kind of knowledge “knowledge in intention.” What makes the proposal so natural is that it explains why we not only don’t have to “listen” to what we are saying in inner speech but don’t need any other method of finding out about the content of what we are saying either: the standpoint we occupy in relation to our own actions is that of practical rather than theoretical reason. You might say that much of our thinking is spontaneous. (It wouldn’t be much use if it weren’t.) It would be quite mistaken, though, to assume that spontaneous actions cannot be intentional, at least in a broad sense of “intentional.” They may involve spontaneously acquired intentions rather than premeditated ones. So they may not be intentional if that is taken to imply prior intent. But they can still be voluntary actions, performed “knowingly” (albeit without foreknowledge), in virtue of one’s possession of “knowledge in intention.” Intentions, though, have a first-personal content, and could not play their distinctive causal and rational role without having a first-personal content. So your knowledge “in intention” of the content of an act of inner speech is knowledge of what you yourself are imagining saying. If the act constitutes an episode of thinking, you know what it is you are thinking. If introspective awareness of an episode of thinking that involves “knowledge in intention” of the content of an act of inner

---

5 For an exception, see Byrne (2011). Central to Byrne’s account are two assumptions: “One can know what one thinks by hearing one’s outer speech. Likewise, one can know what one thinks by “hearing” one’s inner speech” (2011, p. 155). Byrne assumes, further, that “inwardly speaking about x” is a matter of “auditorily imagining words that are about x.” What these assumptions have in common, in my view, is that they misrepresent us as observers of our own current mental activities.

6 See Anscombe (1957). The concept of “knowledge in intention” can and should arguably be detached from the theoretical elaboration it receives in Anscombe’s hands. See Wilson (2000) and Roessler (2010) for discussion.

7 For illuminating discussion of this point, see Campbell (1994, chapter 5).
speech that p, then such introspective awareness just is awareness of oneself silently saying, and thinking, that p. Transparency would be vindicated.

Admittedly my discussion of the inner speech model has been somewhat schematic. Real thinking is a more multifarious matter than imagining saying things. For example, it often involves imagining conversations with others. Sometimes it is far from clear whether one imagines saying something or imagines listening to someone else saying something. I doubt, though, that a more life-like analysis would make a difference to the point that matters in the current context. For even if we sometimes think in the guise of others, or indeed in an indeterminate guise, we know what is being thought in virtue of doing the imagining involved in the thinking. Advocates of the inner speech model may seek to devise an alternative account of introspective knowledge of the content of an episode of thinking, on which such knowledge need not be first-personal. I've argued that appeal to sensory phenomenology would be insufficient. It might be said that the explanation should instead appeal to non-sensory, “cognitive” phenomenology. There is no space here to pursue this, but I think the basic problem is that no method of “finding out,” whether it be sensory or otherwise, can deliver knowledge “from within” of what is being thought.

I would like to end this section by drawing attention to an alternative proposal Langland-Hassan makes in the paper already cited. He suggests that “some reports of voices [i.e., auditory-verbal hallucinations] may be reports of the very same phenomenon that others report as inserted thoughts” (2008, p. 373, my emphasis). The idea is that thought insertion may reflect an unusual auditory phenomenology that does “not fall neatly into any preexisting category” (p. 373). It seems to me that intuitively this is a more promising way to try to secure an explanatory role for patients' auditory phenomenology. But note that the proposal amounts to a wholesale revision of the standard view of thought insertion. It denies that the delusion arises from ordinary introspective awareness of episodes of thinking. If this denial is correct, our inconsistent triad would be a less serious problem than the standard view makes out. The obvious solution would be to reject Alienation. There would be no need to develop a two-concept theory of thought ownership, or to modify Transparency. The weakness of the proposal, though, is that it remains mysterious why (some) patients should describe a certain kind of auditory phenomenology in terms of thoughts, rather than in terms of voices.

**Making Sense**

Thoughts can be judged in some sense “alien” if they are not intelligible in terms of one’s concerns, beliefs, and values. Stephens and Graham think that if a thought were alien, it would, in one sense, not be one’s own thought. They compare this with the case of a wayward bodily movement (2000, pp. 153–154). Suppose you feel your arm going up without any sense of raising your arm. You would presumably acknowledge that your arm went up but quite intelligibly, and perhaps correctly, deny that you raised it. (Patients with anarchic hand syndrome find themselves in just this predicament.)

But what would be the analog of the proprioceptive experience of one’s arm going up? Suppose advocates of the explanatory model accept my suggestion of the previous section. Suppose they grant that to encounter a thought in one’s own mind involves “agentive”
self-awareness in the minimal sense that it involves an awareness of (oneself) saying something in inner speech. Still, they might insist that this anemic sense of “agentive” (i.e., introspective) ownership contrasts with the more full-blooded sense in which I am the agent of only those episodes of thinking that are intelligible in terms of my concerns, beliefs, and values. Would this be a stable combination of views? I think not. If thinking involves “knowledge in intention” of what one is saying in inner speech, finding oneself saddled with an alien, unintelligible thought would not be akin to feeling one’s arm going up. It would more be like finding oneself raising one’s arm, without any sense of why one might have done this. For one thing, one would certainly be aware of oneself thinking something, rather than merely of an episode of thinking occurring in one’s mind. For another, one would be aware not just of the occurrence of an event but of having an attitude—the intention to say something in inner speech—that finds expression in the event. Indeed one would presumably be aware of the attitude as one’s own, supposing that one’s own first-person intentions are easily apprehended as one’s own.

Now it seems to me that as a matter of fact this account is entirely consonant with the way we ordinarily think about thoughts judged to be “alien.” The cellist Pablo Casals reports that when his left hand was badly injured during a mountaineering trip on Mount Tamalpais, his first thought was: “Thank goodness, I’ll never have to play again.” The datum, for Casals, was that he found himself thinking this thought, or equivalently, that the thought occurred to him. The question whether the thought could be understood in terms of his attitudes was not, for Casals, a question that had any bearing on the ownership of the episode of thinking. His own later interpretation was that the thought probably reflected his long struggle with stage fright. But he did not take this to mean that the thought was probably his own. The example also highlights a second problem facing the explanatory model. We are normally aware of our thinking as our own concurrently and effortlessly, without engaging in the construction of an autobiographical narrative that might render the occurrence of the thought intelligible. Stephens and Graham’s version of the explanatory model represents awareness of ownership as an implausibly high-level, reflective matter.

8 Importantly, acts of inner speech that constitute episodes of thinking are informed by further intentions, such as, perhaps, the intention to express a view worth considering. (See footnote 4.)

9 See Gerrans (2001) for helpful discussion of this problem. In Campbell’s theory, the task of determining whether an episode of thinking reflects the subject’s “long-standing dispositional states” devolves upon a subpersonal mechanism whose operation is said to result in an occurrent “sense of ownership.” This is hard to square with the possibility of cases where finding oneself thinking something presents one with an interpretative puzzle, as illustrated in the Casals example. There is also a more basic problem. Why should the question whether an episode of thinking is caused by one’s “long-standing dispositional states” have any bearing on one’s ownership of the episode? Campbell’s answer turns on the idea that the owner, and “agent,” of a token thought is the person “producing” or “forming” it or “bringing it into existence” (see especially Campbell 2002), where this, in turn, is thought to be a matter of the token thoughts being produced by the person’s long-standing dispositional states. Now “being the effect of a propositional attitude” is not sufficient for “being something that has an agent.” (A headache may be brought on by a complex combination of propositional attitudes but it is not clear what it would mean to be the agent of a headache.) If a “token thought” is to be something that has an agent, the attitudes producing it would presumably have to include an intention to produce a (certain kind of) “token thought.” But in that case, ownership could be secured simply in virtue of having a proximal intention to think something now. Intelligibility in terms of one’s “long-standing dispositional states” would be quite irrelevant to the issue of ownership.
Explanatory theorists, then, should reject the view that thinking involves an awareness of saying things in inner speech. They need a conception of thinking on which there can be such a thing as a primitive, wholly “non-agentive” introspective encounter with an episode of thinking in one’s own mind—somewhat analogous to a proprioceptive experience of one’s arm going up. The crucial disanalogy, it seems to me, is that in the introspective case, there is awareness of the intentional content of the episode. The content of the episode is determined by the attitudes finding expression in that episode. So it is hard to see how one can be aware “from within” of the content of the episode, without enjoying an awareness “from within” of the attitudes expressed in it.

**Empirical Versus Autistic-Schizophrenic Delusions**

If Transparency is resistant to revision, the natural course to adopt is to probe Intelligibility. This, according to some psychiatrists, is what we should have done in the first place. The idea that schizophrenic delusions are to be interpreted as rational descriptions of abnormal experiences, according to these psychiatrists, falsifies the nature and depth of the disturbance of the normal state of consciousness in schizophrenia. I want to look at a recent elaboration of this view, from Josef Parnas. I then state what I take to be the two major challenges confronting the task of revising Intelligibility. In the final section I argue that these challenges can be met by taking into account the prehistory of the delusion.

Parnas’s view of what he calls “autistic-schizophrenic” (or “bizarre”) delusions is best understood by way of the contrast he draws with “empirical” delusions. Empirical delusions generally “comply with a normative view of natural causality” (give or take some idiosyncrasies), they tend to be about everyday matters (such as that one’s home is infested with insects), and they reflect concerns that often “caricature human desires familiar to most of us, e.g., becoming rich, starting a software company in order to compete Microsoft out of business” and so forth (Parnas 2004, p. 155). Furthermore, empirical delusions are like ordinary beliefs in being “attached to the intersubjective world”: patients frequently and sometimes passionately try to convince others of the correctness of their delusions. It’s not unusual for them to produce supporting evidence during consultations with their psychiatrist. In contrast, not only is the content of autistic-schizophrenic delusions far removed from everyday concerns (as thought insertion illustrates), but they are “intersubjectively disengaged,” partly in that patients tend not to argue for their delusions. Indeed they can be puzzled by the request for supporting reasons. To a patient with schizophrenia, Parnas suggests, a delusion is given with the same kind of “experiential evidence” we would normally associate with avowals of pain (2004, p. 158). When asked about their reasons, these patients may feel the way most of us would feel if asked “What makes you think you have a headache?”

In the case of empirical delusions, there can be no doubt that patients know what they are talking about. The patient who thought her home was infested with insects, for example, was remarkably resourceful in explaining away unwelcome evidence. Still, her conception of what evidence was relevant to her claim was intact, reflecting her undiminished
grasp of the content of her belief. In the case of autistic-schizophrenic delusions, the question whether patients really know what they are saying is at least a natural question to ask. For their conception of what counts as relevant evidence appears to bear no intelligible relation to the content of the delusion. It certainly seems to bear no such relation in cases where patients are mystified by the very idea that their claims might stand in need of supporting evidence. But even in cases where justifying reasons are provided, patients’ views of what counts as a good reason reflect at best an idiosyncratic understanding of what they are saying.\(^ {10} \)

Parnas’s view may be put by saying that in the light of this disconnection between evidence and content we should abandon the quixotic project of making rational sense of the delusion. To pursue that project is to falsify the phenomenon. It is to mistake an autistic-schizophrenic delusion for an empirical delusion. That is not to say that such delusions are ununderstandable. On Parnas’s view, schizophrenia is a “disorder of consciousness.” It involves a fundamental transformation and disturbance of the normal state of consciousness, afflicting both the “presence” of the world in perceptual experience and patients’ awareness of themselves and others as subjects of experience and action.\(^ {11} \) This “phenomenological” interpretation, Parnas argues, secures “a degree of coherence, and, in contrast to Jaspers’ view, a certain understandability of the psychotic symptoms” (Parnas et al. 2002, p. 135).

But can the phenomenological interpretation furnish an explanation of why patients come to (e.g.,) believe in thought insertion? For advocates of the standard view, this question is critical. Thought insertion is a delusion shared by numerous patients with schizophrenia. This can be no coincidence. The way to establish that it is not is to discover some common causal factor accounting for patients’ acquisition of the delusion. At this point, the “phenomenological” interpretation faces a dilemma, corresponding to two ways of thinking about the explanatory relation between the altered state of consciousness and individual psychotic symptoms. On the first model, we should think of the altered state of consciousness as constituted by the symptoms. This would not necessarily deprive the state of any explanatory role. For, as Brian O’Shaughnessy (another thinker steeped in the phenomenological tradition) has argued, a state of consciousness and its constituent properties may be mutually dependent. The state may be constituted by the co-presence of the properties; at the same time the constituents may depend on the state of consciousness they help to constitute. To illustrate, O’Shaughnessy maintains that being disposed to attend to perceived objects or events is a constitutive element of the state of waking consciousness. At the same time, he argues that the ability to exercise that disposition depends on the presence of the other (mutually dependent) constituents of the state of waking consciousness, such as “occurrent rationality,” self-knowledge, intentional activity, and so on.\(^ {12} \) There is a sense, on this account, in which being in the state of waking consciousness explains one’s possession of the individual abilities that co-constitute the state. Note, though, that the account owes whatever explanatory force it has to O’Shaughnessy’s careful

\(^{10}\) Compare Parnas’s example of a patient who sought to demonstrate that he was Jesus Christ by appeal to these facts: “he felt himself a kind of anonymous being—he did not feel that any predicates really applied to his void,” and “Jesus was also without predicates, because God cannot be said to have mundane features or predicates” (2004, p. 159).

\(^{11}\) For a particularly helpful formulation of the view, see Parnas and Sass (2001).

\(^{12}\) See O’Shaughnessy (2000, chapters 2 and 3).
and detailed analyses of the ways in which the co-constituents of the state of consciousness depend on each other. Parnas and Sass offer no analogous account of any putative explanatory relations amongst psychotic symptoms—say, of the ways in which thought insertion might depend on the altered “presence of the world” in perceptual experience. In any case, a natural thought here is that this whole model looks promising only in connection with psychological abilities or dispositions, rather than particular beliefs. This brings us to the second conception of the explanatory relation between the altered state of consciousness and psychotic symptoms. One might think of a state of consciousness as something that affects the way we perceive things. For example, if you are in a blind rage, you may perceive the recalcitrance of a gadget as a provocation; hence as a good reason to hit it. On this model, though, the state of consciousness affects what we do or believe by affecting our conception of what we have reason to do or believe. Appeal to the state of consciousness would work in tandem with a reason-giving explanation, rather than providing an alternative style of explanation.

We can call this the explanatory challenge. The worry is that the phenomenological interpretation renders delusions inexplicable. If we are interested in the reason why patients believe that thoughts are being inserted in their minds, we really have no alternative but to look into patients’ reasons for holding that belief. Appeal to an altered state of consciousness cannot replace this style of explanation. Nor does it seem plausible that a non-psychological explanation—e.g., in terms of biological factors—alone could adequately account for the acquisition of a belief with a particular kind of content.

Discarding Intelligibility is problematic for another, deeper reason. If reports of delusions are meaningful statements, patients must grasp the meaning of the terms they are using. But grasp of the meaning of a term is inseparable from the disposition to use the term rationally. This raises the question whether autistic-schizophrenic delusions, in Parnas’s sense, are possible. On the face of it, they are not. If there is a real disconnection between the content of the delusion and patients’ conception of the relevant evidence this would simply mean that patients have lost their grip on the meaning of the terms they are using. This is what John Campbell has described as “the basic philosophical problem raised by delusions.” The problem is that “since we have to ascribe meaning in such a way as to make the subject rational, we end up having no way in which to formulate the content of the subject’s delusion.” We face the prospect of having to think of the delusion as “empty speech’ masquerading as belief” (Campbell 2001, p. 91).

It is to avert this unacceptable consequence that the standard view insists that an apparent disconnection between content and evidence must reflect an insufficiently sensitive interpretation. For example, it may initially strike us as unintelligible how inner experience, however unusual, could be taken to suggest that someone else is controlling the thinking going on in one’s head. But once we learn to distinguish between “introspective” and “agentive” ownership, we realize that there is a perfectly intelligible justificatory link between a certain abnormal inner experience and at least one element of the content of the delusion. In the end we have no choice but to interpret apparently bizarre delusions as disguised empirical delusions.

In view of the two challenges, explanatory and semantic, I think simply discarding Intelligibility is not an option. But I want to argue that contrary to the standard view, Intelligibility can and should be modified, along these lines: reports of thought insertion express attitudes that originate in rationally intelligible, coherent beliefs.
Thought Insertion and its History

The preoccupation of patients with schizophrenia with the themes that form the subject matter of their delusions does not suddenly spring into existence with the onset of psychosis. Understanding a delusion such as thought insertion may require charting its development from some non-delusional “precursor.” As Parnas and Sass remark, “one cannot comprehend the delusional transformation in schizophrenia unless the subtler, fundamental features, predating the onset of psychosis, are also taken into account” (2001, p. 102). They provide the following richly suggestive example in which a patient describes his state of mind during the prodromal phase of schizophrenia:

I bypass a window display of a shop in which there are exposed bicycles and bicycle parts; [in a wheel] all the metal spokes cross each other in sharp angles before they reach the axle . . .  the axle turns around with the spokes. No, it is not the axle that rotates; it is the bar, a piece of steel. The axle does not exist; it is just a mathematical line, perpendicular to the plane of the wheel that is determined by the spokes, by forty straight lines. However, this is not necessary either: Only two lines are needed to determine a flat surface. And the circumference? $2\pi r$ is the expression for the length of the felloe, or more precisely, for the theoretical circumference outlines by this inexact circle (i.e., the felloe). Are we able to conceive of an ideal line by paying attention to the lines in nature? Is Spencer’s claim that mathematics originates from experience and induction correct? . . .  These associations . . .  would not seem to me as sick if I were able to master them, like someone who calmly reflects on the matters he is working with, contemplating some professional problems. But when I am thinking in this way, without being able to stop it . . .  I have no mastery over the course of these ideas . . .  it seems to me as if it is not me who generates them. (Hesnard 1909, p. 146; translation and italics by Parnas and Sass 2001, p. 108)

The patient does not believe that he is not generating the ideas troubling him, let alone that they are generated by someone else. Nor is it plausible to interpret him as putting forward the claim that he is not generating them as an epistemic possibility—something for which there is good evidence, as when you remark “it looks as if it’s going to rain.” A more illuminating comparison may be with this claim: “it looks as if the sun is sinking into the sea.” One way to interpret this claim is as an implicitly comparative use of “look.” (See Martin 2010.) The way things look is said to be relevantly similar to a look that might lead someone who was ignorant of astronomy to think the sun was something that could, and did, sink into the sea. To make that claim, obviously, is not to say that the sun is something that could conceivably sink into the sea. Similarly the patient’s description may amount to the claim that his state of mind was such as might lead one to think one wasn’t generating the thoughts one was thinking. One doesn’t have to regard it as a genuine possibility that certain ideas occurring to one may not be generated by oneself for it to help describe one’s state of mind.  

I think this point may help to illuminate what is surely the most puzzling feature of thought insertion, the nature of the causal relation invoked in the delusion. Thoughts are said to be inserted or “put” into one’s mind by others, rather as if they were physical objects.

13 Parnas and Sass speak of a description of the altered state of consciousness at the “as if” metaphorical level” (2001, p. 109).
The two-concept view holds that this peculiar causal judgment represents, or at least arises from, an unusual application of an entirely commonsensical idea—the idea that a (token) thought is something that has an agent. On this interpretation, it would be natural to expect that prior to the onset of psychosis and delusion, patients might describe their state of mind by saying “when I am thinking in this way it seems to me as if it is not I who is thinking.” It is admittedly hazardous to try to reach general conclusions from a single case. But at least the example quoted by Parnas and Sass points toward a suggestive alternative interpretation. The patient says “it seems to me as if it is not me who generates these ideas.” On the face of it, the notion of ideas being generated—whether by oneself or anyone else for that matter—is not one that forms part of ordinary commonsense psychological explanatory practice. The question of who is generating the ideas that occupy me is not one we ordinarily face, or would find it easy to make sense of. Patients’ concern with what (or who) is “generating” their thoughts may have its origin not in the familiar notion that thoughts have thinkers but in a notion that is as unfamiliar as the state of mind it is intended to articulate—a state of mind in which it seems to one that one’s thinking resists “mastery” and lacks its normal intelligibility, being a matter of mere “associations,” and in which, therefore, it is for one as if one’s thoughts or ideas were something that had an utterly perplexing causal explanation and, perhaps connectedly, were something rather like a physical thing.  

At this point let’s return to the two challenges of the previous section. On the current proposal, the task of explaining thought insertion has to be divided into two parts or stages: an account of the acquisition of the “precursor” belief and an explanation of its transformation into a delusion. This analysis promises to meet the explanatory challenge by tracing the content of the delusion, at least in part, to the perfectly intelligible way in which patients are disposed to describe their state of mind before the onset of the psychosis. This would account for the fact that thought insertion is a delusion shared by numerous patients. It would even be true that the explanation of that fact requires an element of reason-giving explanation, as defenders of Intelligibility insist. But in addition, the explanation would appeal to the “delusional transformation” of the belief. It claims that with the onset of psychosis patients’ altered state of consciousness leads them to take at face value the idea of an utterly unusual, quasi-mechanical explanation of their thinking, previously invoked within the scope of “it seems to me as if.” The disconnection between content and evidence, highlighted earlier, may reflect patients’ tendency to adopt the same posture toward the content of the delusion as, previously, toward the content of the “precursor” belief, treating it as a matter on which they enjoy first-person authority.

But given that disconnection between content and evidence, are we not forced to conclude that patients no longer understand the terms they are using? How can we interpret reports of the delusion as expressive of a real belief when it seems doubtful that patients grasp the content of that belief?

Parnas sometimes characterizes autistic-schizophrenic delusions as “metaphorical thematizations” of patients’ altered state of consciousness (2004, p. 160). One might take this

---

14 This would suggest that a “historical” analysis of thought insertion should modify not just Intelligibility but Alienation as well. The theme of the stream of consciousness assuming physical characteristics is familiar from other schizophrenic delusions, such as “thought pressure.” One patient reported, prior to the onset of psychosis, feeling as if “his consciousness consisted of multiple emanating sources, [. . .] each ‘pulsating’ at its own pace” (Parnas and Sass 2001, p. 108).
to mean that even during the psychotic phase patients are perfectly aware that what they are saying is not literally true, and merely intend it to convey a sense of the strangeness of their state of consciousness. So the delusion would turn out to be a coherent, rationally intelligible belief after all. Now there is some support for the idea that, “at some level,” patients do retain a proper sense of reality. For example, Parnas mentions the “double bookkeeping” characteristic of schizophrenic delusions, the fact that patients’ intentional activities are not always influenced by their delusions. Still, I doubt that a straightforward “metaphorical” interpretation is what Parnas has in mind. I take it “metaphorical thematization” is meant to suggest that the delusion develops out of a (broadly) metaphorical description of patients’ experience. The canonical statement of Parnas’s view may be that bizarre delusions are “distorted metaphors” (Cermolacce et al. 2010, p. 10, my emphasis).

A more plausible response to the semantic challenge is to acknowledge that patients’ conceptual abilities are impaired, but to suggest that impaired conceptual abilities can sustain impaired beliefs—viz., delusions. The standard theorist’s bugbear, that unless Intelligibility is correct reports of delusions turn out to be “empty speech,” cannot be taken seriously. Or rather, it poses a serious threat only if we accept a certain kind of instrumentalism about propositional attitudes. If the correctness of an attribution of a belief consists in its usefulness in the context of the enterprise of rational prediction and explanation, the very idea of a belief that is impaired, in that its basis and consequences are somewhat out of step with its content, makes no sense. But that is not how we ordinarily think of propositional attitudes. The idea that an attitude may fail to play its proper rational-explanatory role, due to the subject’s state of consciousness, is arguably one that is familiar to commonsense psychology. True, autistic-schizophrenic delusions involve something more serious than a temporary failure to make proper use of one’s intellectual powers. But there are two reasons to think that patients’ conceptual abilities have not faded away completely. One is that even during the psychotic phase all their propositional attitudes are not suddenly turned into delusions. As Louis Sass observes, “it is remarkable to what extent even the most disturbed schizophrenics may retain, even at the height of their psychotic periods, a quite accurate sense of what would generally be considered to be their objective or actual circumstances” (1994, p. 21). The force of the second reason depends on the intuitive plausibility of a “historical” account of thought insertion. It seems compelling that there is some continuity between the content of the “precursor” belief and that of the delusion—that the latter is a distortion of the former. And that continuity implies the persistence of the relevant conceptual abilities. In O’Shaughnessy’s words, “insanity arises upon the ground of sanity: an inherently impossible cognitive representation of the world must be a distortion of an already acquired viable representation” (2000, p. 150).

Acknowledgments

Previous versions of this chapter were presented at workshops and seminars in Copenhagen, Birmingham, London, and Warwick. I’m very grateful to participants for their suggestions

15 O’Shaughnessy’s searching analysis of the state of drunkenness provides an illustration (see O’Shaughnessy 2000, chapter 3).
and criticism. I’d also like to thank Richard Gipps for extremely helpful comments on the penultimate draft.

References