Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations

Simon R. Jones *, Charles Fernyhough

Department of Psychology, University of Durham, South Road, Durham, DH1 3LE, UK

Received 8 November 2005
Available online 7 February 2006

Abstract

Passivity experiences in schizophrenia are thought to be due to a failure in a neurocognitive action self-monitoring system (NASS). Drawing on the assumption that inner speech is a form of action, a recent model of auditory verbal hallucinations (AVHs) has proposed that AVHs can be explained by a failure in the NASS. In this article, we offer an alternative application of the NASS to AVHs, with separate mechanisms creating the emotion of self-as-agent and other-as-agent. We defend the assumption that inner speech can be considered as a form of action, and show how a number of previous criticisms of applying the NASS to AVHs can be refuted. This is achieved in part through taking a Vygotskian developmental perspective on inner speech. It is suggested that more research into the nature and development of inner speech is needed to further our understanding of AVHs.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Agency; Auditory verbal hallucinations; Forward model; Inner speech; Vygotsky

1. Introduction

The phenomenon of auditory verbal hallucinations (AVHs), where individuals report hearing speech in the absence of any external stimulation, continues to puzzle psychiatrists and psychologists. Schneider (1959) classified AVHs as a first-rank symptom of schizophrenia, reflecting the approximately 60–74% of those with schizophrenia who report experiencing them (Slade & Bentall, 1988; Wing, Cooper, & Sartorius, 1974). However, a movement has developed away from understanding AVHs as necessarily signifying pathology, and towards an acceptance that voice-hearing can be a part of normal experience (Johns & van Os, 2001). Furthermore, there do not seem to be radical differences in the structure and functions of AVHs between voice-hearers with a diagnosis of schizophrenia and those without (Leudar, Thomas, McNally, & Glinski, 1997). Whether in a clinical or non-clinical sample, one of the fundamental characteristics of AVHs is their alien quality. In this...
article, we take a new look at the question of how it is possible that a self-generated cognition may come to be experienced as produced and performed by an agent other than the self.

2. Explaining agency: Is it me?

Frith and colleagues (e.g., Frith, Blakemore, & Wolpert, 2000) have developed an elegant model of the passivity experiences, such as delusions of control, found in schizophrenia. This model attributes such experiences to deficits in a postulated neurocognitive action self-monitoring system (NASS), and has had its predictions supported by empirical research (e.g., Blakemore, Wolpert, & Frith, 1998). The NASS model is based on Miall and Wolpert’s (1995) forward model which was developed to model systems in which, due to temporal constraints, it makes sense to base decisions on the predicted consequences of actions. Frith and colleagues’ utilization of these ideas may be summarized as follows (adapted from Blakemore, 2003; Frith et al., 2000; see Fig. 1).

First, a representation is created of what motor command is needed to achieve a particular goal, based on the estimated current state of the system and the desired end-state. The motor command needed to achieve this goal is then issued. In parallel to this an efference copy of the motor command is also issued. The efference copy is used by the brain, in conjunction with knowledge of the current state of the system, to create a prediction of what will happen if this motor plan is executed. It is proposed that if the actual sensory feedback matches the predicted state then awareness of initiation of movement will remain based on the predicted state. In this scenario, awareness of performing a motor action is hence based on the predicted state, which is available before the movement is actually performed. This results in individuals being aware of the occurrence of their motor action around 50 to 100 ms before they have actually moved (Haggard, Newman, & Magno, 1999; Libet, Gleason, Wright, & Pearl, 1983). Thus in Fig. 1, awareness of motor actions (although not self-authorship) occurs at the time of predicted state generation, which temporally precedes actual performance of the action. If the action is self-produced then predicted sensory feedback should be cancelled out by reafference from the actual sensory feedback. If this occurs then there is perceptual sensory attenuation of the motor act, meaning that one does not feel or pay as much attention to the movement. If the actual movement does

Fig. 1. Forward model of motor control (adapted from Frith et al., 2000).
not match the predicted movement, due for example to a defective predicted state mechanism, then the predicted sensory feedback and actual sensory feedback signals will not cancel each other out.

Frith et al. (2000) use this postulated mechanism to explain why an action may be actively experienced as performed by the self or passively experienced as performed by an alien ‘other.’ Blakemore (2003) has detailed the mechanism through which we come to experience an action as authored by another. She claims that it is the predicted sensory feedback not matching and consequently not canceling out the actual feedback, leading to greater activity in the parietal cortex, which makes the movement feel “externally controlled” (p. 651). Support for this assertion comes from the work of Spence et al. (1997), who suggest that over-activity of the parietal cortex may contribute to the feeling that willed actions are externally controlled in patients suffering from delusions of control.

The mechanism through which we come to experience an action as authored by ourselves has been detailed by Frith (2002). Frith claims that we get our awareness of authoring movements before the comparison has been made between the predicted and actual feedback. To argue for this, Frith draws on Wegner and Wheatley’s (1999) work showing that the temporal contiguity of a thought of what is about to happen (e.g., hearing the word ‘swan’) followed by it actually happening (e.g., finding that your hand has just moved a pointer to a swan) causes the emotion of self-authorship. Wegner (2002) calls this “apparent mental causation” (p. 64). Frith applies this to the forward model by suggesting that the “emotion” (Wegner, 2002, p. 325) of self-authorship is created when awareness of an action about to occur, based on the predicted state (available 50–100 ms before we move as discussed above), is promptly followed by the actual action.

The forward model can be applied to deviations from the normal processes of action authorship in the following way. If the predicted state mechanism is malfunctioning, either through efference copy information not reaching it or through some other impairment, then first Wegner’s mechanism of apparent mental causation cannot work, meaning that the actor does not feel the authorship emotion, even though the action is self-initiated. Second, the high level of parietal cortex activation (due to non-cancellation of predicted and actual feedback) is the same as if the movement were passive (i.e., caused by someone else). This gives the event the same ‘feel’ as a passive or externally caused action, and the actor hence feels as though someone else caused the action.

3. The application of the forward model to AVHs

Somewhat surprisingly Frith and colleagues have only applied this version of their forward model to abnormalities involving overt actions such as delusions of control and anarchic hand (e.g., Blakemore & Frith, 2003; Frith, 2002; Frith et al., 2000). Frith and colleagues have not attempted to use it to explain phenomena not involving overt movements, such as AVHs. One reason for this may be criticisms (e.g., Gallagher, 2004; Stephens & Graham, 2000) of Frith’s previous attempt to apply his prior model to cognitions (Frith, 1992). We address these criticisms, and offer some possible responses to them, in Section 4 of this article.

Frith et al.’s (2000) model has, however, recently been applied to AVHs by Seal, Aleman, and McGuire (2004). Seal et al. begin by noting that any neurocognitive model of AVHs needs to account for how self-generated thought, misperceived as speech, is experienced as unintended. The highlighting of unintendedness as a defining feature of AVHs leads Seal et al. to apply the forward model to AVHs in a particular manner. Their application of the forward model to AVHs is founded on the assumption that inner speech is the primary material of AVHs. The involvement of inner speech in AVHs is accepted by many psychologists (e.g., Bentall, 2003; Fernyhough, 2004), and supported by empirical research such as Gould’s (1950) study showing that when patients hallucinated, subvocalizations occurred which could be picked up with a throat microphone. That these subvocalizations were causally responsible for the inner speech perceived in AVHs, and not just echoing it, was suggested by the work of Bick and Kinsbourne (1987), who demonstrated that if people experiencing hallucinations opened their mouths wide, stopping vocalizations, then the majority of AVHs stopped.

Having made the assumption that inner speech is the primary material of AVHs, Seal et al. next propose that, “once some trigger event brings about the generation of the AVHs motor commands are issued and inner speech is produced” (p. 65). This makes the (unexamined) assumption that inner speech may be conceptualized as an action, and that it is therefore valid to apply the forward model to inner speech. Once this assumption has been made, Seal et al. are then able to apply Frith et al.’s (2000) forward model to the phenomenon of
inner speech, postulating that the experience of unintendedness, characteristic of AVHs, is due to the predicted sensory consequences of inner speech initiation being distorted or absent. This distortion or absence of the predicted state means that, when the actual sensory consequences of the inner speech command occur, there is no predicted state to accurately attenuate it. As Seal et al. (2004) phrase it, “The experience of unintendedness can be accounted for by failure of feed forward information” (p. 64). This concept of the unintendedness of inner speech is part of Seal et al.’s proposal that the failure of the predicted state leads to inner speech whose origins (self or other) is experienced in “various states of ambiguity…rather than two distinct states of awareness (self/other)” (p. 65). The proposal is then made that top-down factors, such as attributional biases, lead to the unintended inner speech being experienced as other-generated. This mechanism is illustrated in Fig. 2.

We suggest, however, that this application of the forward model to inner speech is problematic. Contrary to Seal et al.’s assertions, the forward model outlined by Frith et al. does not state that the absence (or distortion) of the predicted state causes the experience of unintendedness, which is then resolved into a feeling of self or other authorship by preconscious attributions. Instead, as discussed above, the forward model has a two-part mechanism, in which the emotions of self-authorship or other-authorship of an action are determined by the Frith/Wegner and Blakemore mechanisms, respectively.

Our alternative interpretation of how to apply the forward model to inner speech is shown in Fig. 3. First, as in Seal et al.’s model, the brain either produces a degraded predicted state or fails to produce a predicted state at all. The reasons for this are still poorly understood, but it seems likely that it will involve a particular neurological deficit, potentially modulated by stress. The consequences of failing to generate a coherent predicted state from the initial inner speech motor command are likely to be that awareness of performing inner

![Fig. 2. Seal et al.'s (2004) application of the forward model to AVHs.](image-url)
speech cannot occur, as Libet et al.’s (1983) work suggests it would, 50–100 ms before inner speech actually occurs. Consequently, Wegner’s (2002) mechanism of apparent mental causation (responsible for generating the emotion of self-authorship) is unable to operate. Instead, Blakemore’s (2003) mechanism of non-cancellation of predicted and actual feedback, leading to increased parietal cortex activation and hence the feeling of external control, would be operative, leading the event to be attributed to an external cause.

To summarize, Seal et al.’s application of the forward model (shown in Fig. 2) proposes that a mismatch between predicted and actual state leads to an experience of unintended inner speech which is then resolved into self/other authorship by preconscious attributions. Instead, we argue (see Fig. 3) that the failure of the predicted state means that the emotion of self-authorship is not felt, and that instead the mismatch of predicted and actual state leads to the experience of other-authorship of inner speech. This model, contra Seal et al., does not require that there be a feeling of unintendedness, the ambiguity of which must then be resolved into self/other authorship by preconscious attributions.

Fig. 3. Revised application of the forward model to AVHs.
4. Inner speech as a kind of action

One potentially problematic aspect of Seal et al.’s (2004) account is the unexamined assumption that it is possible to transpose Frith et al.’s (2000) model from overt actions to the process of thinking in inner speech. Gallagher (2004) has made a strong case that it is incorrect simply to transpose Frith’s forward model from its successful explanation of passivity experiences, involving overt actions, to cognitive phenomena such as AVHs where there is no overt behavior. Gallagher suggests that any such arguments are based on the erroneous assumption that “thinking is a kind of action” (p. 6).

Gallagher’s suggestion is that Frith and colleagues’ use of the forward model makes sense for overt actions, as there are at least two clear reasons why a predicted state is needed for overt actions. First, we need to know whether our actions are internally caused (i.e., if I lift my arm) or externally caused (i.e., if someone else lifts my arm). Second, if our action is not going to achieve our goal we need to know this in advance so we can make the necessary adjustments. Gallagher goes on to point out that if the forward model is applied to purely internal events such as thoughts, then these two reasons for having the predicted state no longer come into play. First, why would we need to know whether a thought has been caused by us or by an external agent? As Frith and Done (1988) point out, if all our thoughts are internally generated, there is no possibility of our having thoughts other than our own. Stephens and Graham (2000) expand on this by pointing out that whereas we have to distinguish our actions from the actions of others, “we are never confronted with the problem of having to sort out our own thoughts from other people’s thoughts” (p. 138).

However, the issue of how we judge the ownership of our thoughts makes more sense when inner speech and verbal thought are considered within their developmental context. A key figure in this respect is the developmental psychologist, Vygotsky (1934/1987). He suggested that inner speech constitutes a form of verbal self-regulation that is derived from semiotically mediated exchanges with interlocutors in the social world. Dialogue that originally exists on the interpsychological plane, as an exchange between individuals, is reconstructed on the intrapsychological plane as inner speech or verbal thought. Generally speaking, Vygotsky’s hypotheses about the development of inner speech (and its semi-covert precursor, private speech) have been supported by empirical research (Berk, 1992; Winsler, 2004).

The implications of these ideas about inner speech for the present discussion are, first, that thought has a social origin and, second, that the challenges of determining the ownership of one’s utterances continue when those utterances are transferred to the intrapsychological plane. Indeed, a major cognitive challenge for children is to distinguish self-as-speaker from other-as-speaker (Fernyhough & Russell, 1997). Fernyhough (2004) has noted that a Vygotskian approach to inner speech, whereby the experience of alien voices is explained in terms of atypical patterns of internalization of external dialogue, can help to resolve the paradox that AVHs are both experienced as alien and at the same acknowledged to be of the self (Leudar & Thomas, 2000).

Taking seriously the idea of ownership of thoughts would thus militate against Stephens and Graham’s (2000) conclusion that a system that assigns ownership of thoughts is just an “ad hoc” (p. 138) creation devised solely to help explain delusions of thought insertion. Rather, a Vygotskian perspective on inner speech entails an assumption, similar to that made in other recent discussions of the ‘extended mind’ (Clark, 1997, 1998; Clark & Chalmers, 1998), that mental activity is not necessarily co-extensive with the boundaries of the biological organism. For present purposes, the implication of Vygotsky’s ideas about the social origins of thought is that inner speech retains the dialogic quality of the external exchanges from which it derives (Fernyhough, 1996), and that thinking is therefore naturally permeated by other voices (Fernyhough, 2004).

A related question posed by Gallagher (2004) is whether the process of planning and executing an utterance in inner speech can be mistaken in the same way that it can for the generation of overt utterances. In other words, is there an internal, purely cognitive equivalent to the process of making subtle adjustments to an action plan? Although the relevant data for inner speech are necessarily lacking, due to the unobservability of such utterances, some interesting findings from the study of children’s private speech (seen by Vygotsky to represent a way-station between external dialogue and verbal thought) are relevant to this question. Manfra, Tyler, Shiflett, and Winsler (2003) reported evidence that preschool children apply the same correction for speech errors and dysfluencies to their private speech as they do to their social utterances. If private speech is seen as continuous with verbal thought, as the Vygotskian approach entails, it suggests that similar corrections
to inner speech might be plausible. In other words, it seems plausible that individuals can indeed act as though an utterance in inner speech is wrong, and needs adjusting.

A further objection raised by Gallagher's (2004) critique concerns the problem of unbidden thoughts. He notes that thoughts can strike us ‘out of the blue’ without us having a sense of agency, and yet we do not attribute these thoughts to someone else (as a person with psychosis may do). However, there is no evidence that it is these out-of-the-blue thoughts that those diagnosed with psychosis experience as AVHs. Indeed, as Fernyhough (2004) noted, we know very little about the normal processes of inner speech and verbal thought in disorders such as schizophrenia. It may be, conversely, that these out-of-the-blue thoughts are responsible for the AVH-like experiences found in non-clinical populations, but not full-blown AVHs. In the model outlined above, thoughts that occur to us in this way should not be attributed to someone else, as a predicted state is still generated. Asking voice-hearers (both from the non-clinical and patient populations) whether, and under what conditions, they experience such unbidden thoughts will presumably shed further light on the role of predicted state formation in determining the ownership of utterances in inner speech.

One of the benefits of applying Frith’s forward model to inner speech, which Seal et al. do not note, is that this approach has the benefit of being immune to Akins and Dennett’s (1986) criticism of Hoffman’s (1986) influential account of AVHs. Hoffman (1986) proposed that AVHs were due to inner speech that was experienced as unintended, and tried to explain this unintendedness as due to problems with discourse planning. This led to Akins and Dennett’s objection that, if inner speech is intelligently planned by means of discourse plans, then there must be another intelligent entity planning the discourse plans, and so on, leading to an infinite regress. Frith’s forward model sidesteps this problem by not assigning any crucial role in the model to judgments about the unintendedness (or otherwise) of actions. Rather, the model (when applied to AVHs, as outlined here) proposes a direct causal mechanism leading from the malfunction of the predicted state to the experience of inner speech as being of alien origin, in the absence of any competing self-authorship emotion (see Fig. 3). Failure of the predicted state leads to neurological activity associated with passivity experiences, which may indeed lead to inner speech being experienced as unintended. However, the unintendedness of the inner speech is a result of the failure of the predicted state and follows from the passivity experience itself. Thus, unintendedness is a consequence, not a cause, of the AVH and hence does not play the problematic theoretical role that it does in Hoffman’s (1986) model.

A number of further challenges for the present model need to be addressed by future research. One important issue is to pinpoint precise brain areas or networks involved in the generation of (and, in cases of damage, the failure to generate) the predicted state. Leube et al. (2003) have suggested that the neurological instantiation of a deficit in the efference copy mechanism may be located not in specific brain areas, but rather may arise from defective interactions between perceptual and motor areas. These areas are likely to involve the cortical network that de Vignemont and Fourneret (2004) note as being involved in action attribution, which include the prefrontal and the parietal cortex, the supplementary motor area, and the cerebellum. However, current research in this vein has tended to focus on the role of the cerebellum in signaling the sensory discrepancy between the predicted and actual sensory consequences of movements (e.g., Blakemore, Frith, & Wolpert, 2001) and not on the neural correlates of the predicted state itself. Another route of entry into understanding the neural underpinnings of the predicted state mechanism is via neuropsychological syndromes, such as the anarchic hand, which is associated with lesions in known brain regions. In the case of the alien hand syndrome, patients make involuntary yet seemingly purposeful limb movements (Giovannetti, Buxbaum, Biran, & Chatterjee, 2005), suggesting a breakdown in the forward model, but without the feeling that their actions are caused by another. As this syndrome is associated with medial frontal and callosal lesions (Della Sala, Marchetti, & Spinnler, 1991), this identifies these as regions to be potentially incorporated into any neural instantiation of the forward model.

A further challenge, noted by Gallagher (2004), is the selectivity problem, which recognizes the episodic nature of positive symptoms. That is, if the predicted state is malfunctioning, why do all thoughts not seem alien? The first point to note is that the predicted state may only intermittently be malfunctioning. This raises the further question of what is it that causes the predicted state to fail. A possible answer is that it could be the same factors that Bentall (1990) has described as affecting our judgements about whether an event is public or private: namely, factors such as stress-induced arousal, our ability to use cues, perceptual attenuation, or effects of suggestion. These remain areas for future empirical research. The second point relates to the fact,
noted above, that we do not have sufficient understanding of normal inner speech in individuals who experience AVHs. Only a greater understanding of the phenomenology of inner speech in affected individuals will allow us to specify how voice-hearing may be determined by the breakdown of the predicted state mechanism.

References


