Sam Wilkinson
and Charles Fernyhough

Auditory Verbal Hallucinations and Inner Speech

A Predictive Processing Perspective

1. Introduction

Inner speech is a pervasive feature of our conscious lives.[1] But what is inner speech, and what happens in unconscious processing that makes it the conscious experience that it is? A clue to answering this can be found in cases where the mechanisms that produce inner speaking behave unusually. In this paper, we suggest an account of a specific instance of this, namely, a particular subtype of auditory verbal hallucination (AVH), and draw some lessons about the processes that underlie normal inner speech.

An AVH involves, roughly, the experience of hearing a voice in the absence of anyone actually speaking. As a phenomenon, it varies enormously in a number of ways: in how it presents itself phenomenologically, in terms of the context in which it occurs, and arguably in what causes it. This has lead some theorists (Jones, 2010; Wilkinson, 2014; Smailes *et al.*, 2015) to claim that there are subtypes of AVHs, and that these amount to fundamentally different phenomena, underpinned by different mechanisms and different aetiologies. Three identified subtypes are memory-based, inner speech-based and hypervigilance hallucinations. As the names suggest, the 'raw materials' for memory and inner speech-based hallucinations are episodic memories and episodes of inner speech respectively. In contrast, hypervigilance hallucinations involve the moulding and boosting of ambiguous environmental stimuli into voices (as such, they are strictly speaking not so much

[1] At least for most of us; for individual differences see Hurlburt *et al.* (2013).

hallucinations as illusions). Our focus in this paper is on inner speech-based AVHs, and what they tell us about inner speech more generally.[2]

It is worth mentioning that the order of explanatory primacy is normally the reverse of what we are doing here. Theorists tend to use inner speech (which they take to be relatively un-mysterious) to make sense of AVHs (which they take to be relatively mysterious) and not vice versa. However, it seems to us that, despite its prevalence and familiarity, the nature of inner speech is far from self-evident. Given this, it makes sense to start, for the sake of inquiry, with the hypothesis that at least some AVHs are instances of pathological inner speech, and then to ask: what kind of thing must inner speech be in order for it to play this role in the generation of AVHs?

Before moving on, it is important for us to get clear on what kind of thing we are referring to by 'inner speech'. By that term one can be referring either to a particular experience, with a particular phenomenology, or to a particular feature of human cognition, which makes use of particular mechanisms, say, and which sometimes gives rise to that phenomenology, but which needn't always (for example, when it is disrupted in certain ways). In the former sense, the subtype of AVH that interests us is not an instance of inner speech, even though it may be generated by the processes that usually generate inner speech. In the latter sense, that subtype is, or is partly constituted by, an instance of inner speech. We will use the term 'inner speech' in the latter sense, although nothing of substance hangs on this terminological decision, and we acknowledge that both are valid senses of the term 'inner speech'.

2. A Predictive Processing Account of Auditory Verbal Hallucinations

In this section, we present an account of AVH that is built within the predictive processing framework (PPF). Since this account arose in part as a reaction to self-monitoring accounts of AVHs, we begin by presenting these accounts, and then move on to the predictive processing accounts. Then, in Section 3, we will explore the potential for predictive processing accounts of inner speech.

---

[2] Some theorists don't buy into subtypes, but if they adopt inner-speech based accounts across the board, then what we say will be of relevance to them. It is only those who think either that AVHs are homogeneous and nothing to do with inner speech, or who think that there are subtypes, but none of those subtypes are inner speech-based who will take issue with our starting point.

2.1. Self-monitoring accounts of AVH

Self-monitoring accounts are often viewed as unifying accounts of the positive symptoms of schizophrenia.[3] Among these symptoms are delusions of control, AVHs, and thought insertion. What these symptoms all have in common is that they are instances of 'self-monitoring' having gone awry, which roughly means that a 'self-produced' or endogenous phenomenon fails to be recognized as such by the nervous system. These symptoms differ in so far as the phenomenon that is failing to be self-monitored differs. In AVHs and thought insertion, it is often taken to be inner speech. In delusions of control (and experiences of passivity) it is overt bodily action.

So what is this 'self-monitoring'? Perhaps the first theorist to make use of the idea of self-monitoring was Helmholtz (1866). He was not concerned with pathological cognition, but with healthy visual perception. In particular, he wondered, when an image moves across the retina, how does our brain know whether it is the world moving across our eyes or our eyes moving across the world? He suggested that when our eyes move there is a motor command, and that a copy of that motor command, later called an 'efference copy', is used by the brain to calculate a prediction of the sensory consequences of the upcoming eye movement. If this prediction is accurate and the predicted and actual sensory consequences match, then the brain 'infers' that the change was self-generated and the conscious percept is interpreted accordingly as a case of the eye moving across the world. We can experience for ourselves what happens when there is no motor command, and hence no adjustment, when we move our eye directly with our finger: the world itself seems to move, namely, the brain 'thinks' it is the world moving across the eye rather than vice versa.

These ideas were, much later, applied to psychosis (Feinberg, 1978). Although Feinberg's initial paper was on 'thought' (which he took to involve motor mechanisms) and thought insertion, we introduce the self-monitoring account with delusions of control, since it is clear that, if anything involves motor commands, it is overt bodily actions. In delusions of control, a subject claims that somebody else is controlling her actions. Frith and Done (1989) claimed that here there is a mismatch between the predicted and actual sensory consequences of the bodily movement, with the result that (as with Helmholtz's ocular example) the movement is attributed to an external source. In Helmholtz's

---

[3] Needless to say, in reporting these accounts we are remaining silent on the validity of the concept of schizophrenia. For the record, we have doubts that all of those who standardly get the diagnosis of schizophrenia suffer from the same unified condition.

example, the recognition by the nervous system that a certain stimulus is self-produced, due to this matching between the predicted and sensory consequences of movement, causes a correction of the conscious percept. In contrast, in more typical bodily motor control, this matching results in 'sensory attenuation', namely a decrease in the intensity of the sensation. In effect, when there is sensory attenuation, your nervous system is telling you: 'You don't need to pay attention to this: it's only you.'

One striking datum that seems to support the hypothesis that something has gone wrong with this kind of self-monitoring in schizophrenia is the reported finding that subjects with diagnoses of schizophrenia can tickle themselves. The postulated explanation for this is that there is a mismatch between expected and actual sensory consequences and the sensory consequences are not attenuated: the tickling sensation is like being tickled by somebody else (Frith, Blakemore and Wolpert, 2000).

Several theorists (Feinberg, 1978; Frith, 1992; Jones and Fernyhough, 2007; Seal *et al.*, 2004) have attempted to explain AVHs in terms of these same self-monitoring abnormalities operating on inner speech. On these accounts, AVHs are instances of badly monitored, and hence unattenuated and externally attributed, inner speech.

2.2. Problems for the self-monitoring account of AVH

What is wrong with the self-monitoring account of AVH? As Wilkinson (2014) points out, there are potentially problems in accounting for (i) the phenomenology of AVH, and (ii) their variety. The first of these is effectively the issue of how we explain the transformation, in phenomenology, from inner speech to AVHs. The second of these concerns the issue of accounting for the wide varieties in AVHs with one model. This second worry can be overcome simply by saying that only *some* AVHs are misattributed episodes of inner speech arising from self-monitoring abnormalities. This kind of strategy seems like a sensible move regardless of what explanatory model you are trying to promote: AVHs are varied in how they present, in their contexts of occurrence, and in their apparent causes. Whether the first worry can be overcome is still a matter of debate (see, e.g. Cho and Wu's, 2013, attack on inner speech-based approaches and Moseley and Wilkinson's, 2014, defence), but it seems that acknowledging, on the one hand, the complexity and variety of inner speech phenomenology (McCarthy-Jones and Fernyhough, 2011), and the effect of the postulated lack of 'attenuation' resulting from failed self-monitoring, may go some way towards answering this worry.

Another worry may not come from whether the phenomenon to be explained (AVH) seems to fit the account, but rather from the viability of the very idea of monitoring inner speech (regardless of what phenomenon a deficit of such monitoring might generate). First of all, it is not obvious that inner speech involves motoric elements, and, so, where is the motor command that self-monitoring is supposed to exploit? This concern can be addressed, however. Motoric involvement in some forms of inner speech has been empirically supported by electromyographical (EMG) studies, some of which date as far back as the early 1930s (e.g. Jacobsen, 1931). Furthermore, later experiments made the connection between inner speech and AVH, showing that similar muscular activation is involved in healthy inner speech and AVH (Gould, 1948; McGuigan, 1966).

However, demonstrating motoric involvement in both inner speech and (at least some) AVHs doesn't let the self-monitoring theorist off the hook. It is not just motor commands that are important for self-monitoring, it is also the predicted and actual sensory consequence of the monitored phenomenon, and the match or discrepancy between the two. But what *are* the *sensory* consequences of inner speech? Is inner speech sensory at all? If so, where is the stimulus? Furthermore, since it doesn't occur in three-dimensional space, does it even *need* monitoring? These questions point towards a more fundamental worry, namely that the self-monitoring mechanism is not actually very well understood at the neural level. In a related manner, the postulated self-monitoring mechanism seems to be little more than a re-description of the computational task that any active system would need to do in order to distinguish what it does from what is done to it. In contrast, predictive processing accounts start from a general theory of what the brain does, and how this is implemented at the neural level (Friston, 2005). It then turns out that the self-monitoring task that needs to be achieved falls naturally out of this (along with many other tasks besides, e.g. making sense of noisy and ambiguous perceptual inputs).

Indeed, perhaps the main problem with the self-monitoring account actually has less to do with the account itself, and more to do with the overall *framework* within which the account operates, namely, how *cognition generally* is taken to work, how the brain processes information from the outside world and how that relates to conscious experience. Self-monitoring accounts try to explain, within a standard framework for understanding cognition, why someone is having an experience that usually occurs with a particular environmental stimulus (i.e. a speech sound), in the absence of that stimulus. The answer that the self-monitoring account gives is that there *is* a stimulus of sorts, it just hasn't been recognized by the nervous system (it may be so recognized

by the *person,* as when a voice-hearer says 'I know it's just my brain') as a self-produced stimulus. That stimulus is inner speech. But what if this approach is doubly wrong? What if cognition generally, and healthy perceptual cognition, isn't really about the external stimulus in this way? And what if inner speech, more specifically, isn't about, and couldn't be counted as, a stimulus either? We present a general framework, and an account of AVH within it, that pursues precisely this line of questioning.

2.3. From self-monitoring to predictive processing

Some theorists (some of whom were, earlier, the main proponents of the self-monitoring account; compare Frith, 1992, with Fletcher and Frith, 2009) have proposed that the self-monitoring that is taken to go awry in AVH falls naturally out of a basic principle of brain function, namely, *prediction error minimization*. On this account, self-monitoring is not some *additional* aspect of cognition, but is a fundamental part of it (see Pickering and Clark, 2014). One upshot of this is that all of the varieties of AVHs can be accounted for (see Wilkinson, 2014), including those that may not involve motor commands from which forward models could be derived. For example, they can account for the 'hypervigilance' hallucinations (Dodgson and Gordon, 2009) we briefly mentioned in the introduction, in which environmental stimuli are boosted and shaped. This framework for thinking about cognition, and within which self-monitoring emerges from the basic functioning of cognition, is called the predictive processing framework (PPF).[4]

According to the PPF, the brain's main task is to 'infer' from incoming signals what the causes of those signals are. However, the incoming signal underdetermines distal causes: since inputs are noisy and ambiguous, the same stimulation can be brought about by two different distal causes (and different stimulation in different circumstances can be caused by the same distal cause). Given that more than one hypothesis is compatible with the incoming signal, the brain needs to take two things into account: first, the fit of the input with the hypothesis, and, second, how statistically likely that hypothesis is (the 'prior probability'). A hypothesis could fit the input extremely well, but its prior probability could be so low that it isn't even considered. Conversely, an hypothesis could have such a high prior probability, that, even though it doesn't fit the input well, it is settled upon.[5]

---

4 For a fuller presentation of the PPF, see Clark (2013).

5 A nice example of this is the Hollow Mask Illusion. When you are presented with the concave back of a mask, your brain 'corrects' the concave stimulus into a convex stimulus. This is due to the fact that the hypothesis 'convex face' (i.e.

What the selection of an hypothesis does is that it determines a set of predictions about subsequent inputs, namely, inputs that are compatible with the hypothesis. If the hypothesis does a good job of predicting inputs, it is kept. If it does a bad job, it is tweaked or abandoned altogether in favour of another hypothesis that does a better job. In other words, one hypothesis is selected rather than another if it better *minimizes prediction error*.

This picture has interesting consequences for how we are to view the role of input on sensory receptors and its impact on higher cortical regions, and also on conscious experience. According to the PPF, the only information that gets passed on up the cortical hierarchy is *prediction error*. This stands in sharp contrast to the standard view of perception and cognition according to which inputs come in, are processed, and passed on. According to the PPF, what determines your perceptual experience is what your brain has already predicted, your brain's best hypothesis.

This prediction error minimization is not only taken to account for perception and cognition, but for action as well (see, e.g. Adams *et al.*, 2013). Instead of there being motor commands, as on the standard picture, what you have are predictions, which are then fulfilled by the subsequent bodily movement, thereby also being a case of prediction error minimization. This is often called 'active inference', which Pickering and Clark (2014) helpfully gloss as follows: 'the combined mechanism by which perceptual and motor systems conspire to reduce prediction error using the twin strategies of altering predictions to fit the world and altering the world (including the body) to fit the predictions' (p. 1).

Another extremely important aspect of the PPF is that the hypotheses are hierarchically organized, with the hypotheses of one level providing the inputs for the next. 'Hypotheses' can also be talked about in a very 'zoomed out' way, to talk about the overall hypothesis, or in a very 'zoomed in' way, to talk about 'hypotheses' in early stages of perceptual processing. 'Higher' parts of the hierarchy are, roughly, those parts that are further away from the sensory stimulus. These tend to be at longer temporal timescales, and a higher level of abstraction. They might correspond, for example, to the belief that lions are dangerous. 'Lower' parts of the hierarchy are closer to the sensory stimulus. These tend to be operating at shorter timescales, and at low

---

normal face) has a very strong prior probability and that overrides the better fit that the 'concave face' hypothesis has with the incoming signal. This prior probability is generated by the expectation that the faces you will encounter will always be convex.

levels of abstraction. These, for example, might correspond to early stages of perceptual processing: your brain's early statistically-driven attempts to make sense of noisy inputs (see, for example, Gangepain *et al.*, 2012, for strong evidence for predictive processing in auditory word recognition). Of course, in order to express these neurally encoded predictions we need to use rough-and-ready descriptions in natural language (in this case English), but there is nothing linguistic about the priors/hypotheses ('light comes from above'/'This is a face') themselves.

Let's take an example (adapted from Pezzulo, 2014) to illustrate the predictive hierarchy. Suppose that, on the basis of a noise, which you take to be a squeaking window, two hypotheses present themselves about what is going on: either the wind blew the window, or a thief is clambering into your house. At the stage where those two hypotheses are competing, a great deal of ambiguity has already been resolved at lower levels of the hierarchy. For example, in early stages of auditory processing, the qualities of the sound will have been settled upon, giving rise to the conscious experience being a certain way, qualitatively speaking. Higher up the hierarchy, that sound gets interpreted as a creaking window, as opposed to something else. The direction of causation is from the (events represented in the) lower regions of the hierarchy to the (events represented in the) higher regions of the hierarchy. However, the direction of the inference is from the effects to the causes.

One final way in which the framework is made a bit more complex is that, in order to accurately form predictions in a world where the degree of noise varies from context to context, the brain needs also to predict the extent to which it can rely on its predictions. In other words, it needs to form second-order predictions, or estimate the precision of its predictions. In the predictive processing literature, this is called 'precision-weighting', and it amounts to the extent to which prediction error, once generated, is given weight. In contexts of high noise (e.g. in a dark room), the precision-weighting on bottom-up sensory prediction error will be low, and more influence will be placed on top-down influences. That is why at dusk you are more likely to see a tree trunk as a lurking aggressor.

2.4. The PPF and hallucinations

The PPF changes how one thinks of perceptual experience, and, by extension, radically changes one's explanatory focus in trying to account for hallucinations. On a standard framework, where front-line sensory stimuli get gradually processed and passed on up the hierarchy, hallucinations make one wonder, 'Where does this

erroneous sensory stimulus come from?' As mentioned, we can see self-monitoring accounts as making attempts to answer this within a standard framework. Their answer is: they come from the quasi-sensory stimulation of inner speech, which is then misattributed. However, when instead we adopt the PPF, incoming stimuli play a much smaller role in determining the conscious percept, even where veridical perception is concerned. Given that a conscious percept is constituted by the hypothesis that best minimizes prediction error, we don't ask 'Where does the input come from?', since the input alone doesn't (and can't) determine the percept. Rather we ask, 'Why does this hypothesis minimize prediction error?' This general approach makes hallucinations both less perplexing, and less different from veridical perception.

Wilkinson (2014) has suggested at least three different ways in which the hypothesis corresponding to an AVH experience may be selected. These correspond to the three phenomenologically and aetiologically identifiable subtypes mentioned at the outset: inner speech-based, memory-based, and hypervigilance hallucinations. Both inner speech- and memory-based hallucinations are taken to be the result of aberrant weighting on prediction error. In other words, the self-generated hypotheses corresponding to inner speech and episodic memory turn out to generate unexpected levels of prediction error, which results in perception-like hypotheses being selected in an attempt to minimize this. This leads to a 'perceptualization' of the usual experiences of inner speech and episodic memory. In contrast, hypervigilance hallucinations are explained in terms of interroceptive predictive processing (Seth, 2013), where the hypothesis is selected not only based on how well it explains the incoming signal, but on how well it explains both the incoming signal and the subject's interroceptive emotional state. Thus the hypothesis that someone is insulting me explains not just a vague environmental stimulus, but also my state of anxiety and hypervigilance (see Wilkinson, 2014, for more details here).

In Wilkinson (2014), the more original contribution was taken to be the interroceptive account of hypervigilance hallucinations, with existing inner speech (Jones and Fernyhough, 2007) and memory-based (Badcock *et al.*, 2005) accounts merely requiring a slight reframing, from self-monitoring to predictive processing. However, such a reframing is not obviously achieved for either the inner speech or memory subtype, largely because it is not obvious how the PPF accounts for inner speech and episodic memory. In this chapter, we focus on inner speech, although an important area of future theorizing would involve an explanation of how the PPF accounts for episodic recollections.

3. A Predictive Processing Account of Inner Speech

It's all very well saying that AVHs are the result of changes to predictive processing, and that a subtype of AVH involves the mechanisms that are involved in inner speech. But what are the mechanisms involved in standard inner speech? In other words, what does a predictive processing account of inner speech look like?

3.1. Inner speech as 'internalized' outer speech

Before asking ourselves what the PPF makes of a given phenomenon, we need to be clear that we have successfully identified the phenomenon in question. So, what is inner speech, how does it develop, and what purpose does it serve? One very attractive theory, attributed to Lev Vygotsky, which carries both evolutionary and developmental plausibility, is that inner speech starts off as speech, namely, 'overt speech'. That is to say, whatever function inner speech plays, once it has developed, is played by overt speech in children who have not yet developed the capacity to engage in inner speech. This capacity to engage in inner speech is usually seen to involve, at least in part, the capacity to inhibit the overt production of speech (see Alderson-Day and Fernyhough, 2015, for a comprehensive review on the psychology and neuroscience of inner speech).

According to this story, inner speech is the end product of a developmental trajectory that begins with social speech, between an infant and primary caregiver, and then becomes overt private speech, before finally becoming inner speech. 'Private speech' refers to speech that is not produced for the benefit of anyone other than the speaker. Thus, although there is an important sense in which inner speech is always *de facto* private speech, pragmatics dictates that 'private speech' tends to refer to overt private speech, rather than inner speech (since inner speech is obviously private). Young children will first, under the guidance of a caregiver, learn to reason verbally, but out loud, for the benefit of guiding their thinking and attention. Over time, they learn to 'internalize' this speech, or, to phrase it in somewhat less misleading terms, to inhibit its overt production. However, as with many cases of motoric inhibition, vestiges of the motor processes often remain (as clearly seen in Jacobsen, 1931). Furthermore, the reason why an auditory phenomenology is often reported is quite simply because, as with any aborted overt action (motor imagery), the predictions of the sensory consequences of the action come into play, activating sensory (and somatosensory) cortices (this is central to feedback, which is crucial for all successful motoric activity).

3.2. Inner speaking and auditory imagery

What is going on when someone is engaged in inner speech? What constitutes inner speech? It is tempting to think of inner speech in terms of auditory imagery. Engaging in inner speech, on such a view, consists in imagining the sound of you speaking (or imagining hearing yourself speak). There is little doubt that one can imagine the sound of oneself speaking. It is like imagining hearing someone else speak, except that it has the properties of your voice. This, however, is not what inner speech, the phenomenon of primary interest to us, is. As we've seen, inner speech involves not just an auditory/imagistic component, but an articulatory/motoric component, too. Inner speech is agentive and more or less intentional (Jones and Fernyhough, 2007). To the extent that it is correct to speak of inner speech in terms of imagination at all, it does not consist in imagining hearing one's voice: it is the phenomenon of imagining oneself speaking (see Hurlburt, Heavey and Kelsey, 2013, for a phenomenological distinction between 'inner speaking' and 'inner hearing'). In any case, it seems misleading to speak of inner speech in terms of imagination, and here is why.

It is crucially important to differentiate imagination from imagery. Imagination is a personal-level phenomenon: people are engaged in acts of imagination. These acts of imagination enable them to appreciate, in potentially many different ways, non-actual scenarios, and, when they are engaged in such acts, they may be motivated to do so by a number of different things. They may be trying to remember the colour of someone's hair, judge whether they could have jumped over that river, reason about a social situation, or simply engage in imagination for the pleasure of it. These acts of imagination often will recruit or make use of imagery in many modalities, but there will also be aspects to the imaginative experience that aren't imagistic.

Imagery, in contrast, is not a personal-level event. Whereas people imagine things, people don't do imagery. When people imagine things, imagery may be involved. Imagery is also involved in personal-level events that aren't imaginings. For example, imagery may be involved in inner speech, indeed it may even be similar (or even the same imagery) to the imagery involved in imagined speech, but that doesn't make the personal-level act of inner speaking an act of imagining speaking. For a start, with inner speaking, you are not appreciating something non-actual: it is actual. You are speaking.

In short, it is important to understand the relationship between auditory imagery and inner speech, and, in a related manner, to understand that inner speech is not, in virtue of its recruitment of auditory imagery, simply a kind of imagined speech. Two things underpin this;

one is more sophisticated than the other. On the one hand, inner speech involves not just (and sometimes perhaps not even) auditory imagery, but motoric/articulatory imagery as well. In principle, however, there could be imagined speaking that made use of both motoric and auditory imagery, and this leads us on to the second more sophisticated reason why inner speech isn't imagined speech. Inner speech involves making a speech act, involves speaking your mind directly. The fact that someone is engaging in inner speech entails that they are speaking. The fact that someone in engaged in imagining themselves speak not only fails to entail that they are speaking, it actively entails that they are *not* speaking, since they are merely imagining it!

If we are to provide an account of inner speech, we need not only to account for the sensory and motoric imagery that are standardly part of acts of inner speech, but which can also potentially be part of other acts too, but also to account for what distinguishes inner speech from those other acts that make use of similar imagery.

3.3. Motoric and sensory imagery within the PPF

The PPF can very nicely accommodate the aspects of sensory and motoric imagery that are standardly part of inner speech. According to the PPF, all the brain ever does is minimize prediction error. As we've seen, this is taken to account for both perception and action. Whereas in perception hypotheses are selected to generate accurate predictions about the world, thereby minimizing prediction error, in action, predictions are generated which are then to be fulfilled by the action, thereby also minimizing prediction error. As a result, the notion of motor commands, at least as a type of neural activity in their own right, is dispensed with (we could, of course, still call the predictions that bring about actions motor commands—they do, after all, serve precisely that function).

Now, this presents us with an account of 'imagery' for both motoric and sensory domains. Although they are very different, they operate on exactly the same principles, namely, inhibition at a neural level, which within the PPF amounts to down-modulating the weighting/precision of prediction error. This turning down of the gain allows for a decoupling of the brain from the world. It is a way of minimizing prediction error without having to actually match the world (a relatively costly and difficult way, which is why it takes a while for children to master it, and why it is interfered with under conditions of cognitive load). And that is partly, and by definition, what imagery (as opposed to perception) is: something that represents something non-actual. It is a percept or action that isn't actually happening.

How does this relate to self-monitoring accounts? Imagery, both motoric and sensory (of which inner speech is composed), is not a self-produced *stimulus* in need of monitoring. There are no predicted and actual sensory consequences of imagery, where the latter can diverge from the former (as is the case with overt bodily action). Rather, the imagery, like any part of conscious experience, *is the prediction itself*, or, more specifically, a decoupled hypothesis that entails a bunch of deliberately unfulfilled (but prediction-error-minimized, through down-modulation) predictions.

A point of clarification is needed at this point. In this subsection, we have said nothing about inner speech *per se*. We have simply explained how sensory and motoric imagery, both of which seem to be involved in inner speech (as well as many other events besides), are to be viewed within the PPF. What does PPF have to say about inner speech more specifically?

3.4. A predictive processing account of inner speech

As we've said, an episode of inner speech (or, perhaps better, an act of inner speaking) is not an imaginative act. It is not imagining yourself speaking, indeed, it is not imagining anything: it is speaking. But just like overt speaking involves moving your mouth, throat, etc. as well as hearing yourself speak, so does inner speaking, at least often, involve the decoupled versions of these. This amounts to saying that inner speaking makes use of auditory and motoric imagery. So much we've already said. But what makes something an act of inner speaking as opposed to an act of imagined speech is that part of my experience is not only the low-level decoupled hypotheses that determine my imagery (both sensory and motoric), but high-level hypotheses about myself as an agent. Indeed there is a similar distinction when someone else is speaking to me, in a normal overt case, between my low-level hypotheses about sounds (or slightly higher up, phonemes, or higher up still, words, etc.) and my high-level hypotheses about the agent, their intentions, whether these speech sounds constitute a sincere speech act, and, if so, what kind of speech act, and what is the precise communicated content, etc. In inner speech we have all of these hypotheses about ourselves, as we are engaged in inner speech, and, what's more, they are almost always accurate. Verbally reprimanding myself in inner speech involves (i) me actually verbally reprimanding myself and (ii) in so far as I experience that reprimand, my brain having an hypothesis, not just about the words (or phonemes, etc.) used in the reprimand, but about the fact that I am reprimanding myself (which is clearly accurate in this case).

4. Consequences of a Predictive Processing Account of Inner Speech

A predictive processing account of inner speech has a number of interesting consequences. Some of these consequences are shared by compatible but higher-level or developmental accounts, such as Vygotskian accounts. Others are specific to the PPF and the hierarchical arrangement of hypotheses.

4.1. Epistemic consequences

The predictive processing account of inner speech fits nicely with the Vygotskian developmental story, at least in part because they make inner speech and outer speech very similar phenomena. However, this proximity between the inner and outer phenomenon raises an interesting epistemological issue, and it is as follows.

We very often talk to others in order to inform them, either directly or indirectly, of certain things, including states of the world, and our own states of mind. Granted, not all speech acts are informational: they can be imperative, expressive, etc. and the same applies to inner speech. However, in the cases where inner speech is informational, what motivates such speech acts? Why would I bother talking to myself if I already know what I'm going to say? One obvious way out of this problem is to insist that, contrary to our intuitions, we don't really know what we are going to say. Hence our utterances, in inner or outer speech, do not presuppose self-knowledge: they often *generate* it. This conclusion, though arrived at through logical argument, fits extremely well with the PPF. Within the PPF, there is no in-built provision for an introspective mechanism; there is simply the experience of certain percepts, actions, and emotions which all have the potential to feed into higher-level hypotheses I might have about myself.

A related upshot of this is that thinking is in some sense always dialogic. According to the PPF, simple, world-directed cognition involves coming up with accurate hypotheses about the world in the service of the organism's goals. 'Thinking' (among this woolly notion we include reasoning, supposing, wondering) emerges when the organism can decouple itself from the world in the service of goals represented *in absentia*. This involves the generation of things that stand as proxies for the absent (because they are future or distant or abstract) aspects of the world. Speech is a particularly helpful phenomenon that helps us do this (there are likely others), either overtly ('thinking aloud', a phenomenon that literally happens) or in inner speech. In these situations, we are both producers and recipients, and, as such, we are in a constant and inescapable dialogue with ourselves. This,

coupled with the earlier anti-introspectionist epistemological consequence, may even suggest that this dialogicality (although it may not always use the medium of *speech*) is central to the robust self-awareness that humans are capable of.

4.2. The self and other in inner speech

At some relevant level of abstraction, it is clear that we represent other agents, when, among other things, we see them, think about them, hear them talk, etc. Within the PPF, this would correspond to hypotheses at a relevant level in the predictive processing hierarchy (generative models). When someone talks to us, we represent them (we retrieve a previous representation if it's someone we know already, or we use a generic model if it's someone new). It also seems that, in inner speech, we at some level represent ourselves at least implicitly. Now, there is one feature of AVHs, inserted thoughts, and cases of delusions of control that seems somewhat problematic for self-monitoring accounts, and it is that the subject doesn't simply claim to be *passive* in the face of these thoughts, utterances, and actions, but gets a sense that they are the responsibility of another, often quite richly represented, agent (see Wilkinson and Bell, 2015, for a focus on the representation of specific agents in AVHs). This is problematic for the self-monitoring account, because this account only tells us that the subjects ought to experience this as 'not me'. However, they do not explain the move from 'not me' to 'someone else'. Of course, these theorists could say that this is merely an abductive inference based on the feeling of 'not me' ('My actions or inner speech don't feel like me. How do I explain that? It must be someone else').

We note that this inferential step is an under-acknowledged feature of inner-speech models of AVHs. Two retorts to this inferential tactic are that, first, it doesn't explain why that is the explanatory inference so often resorted to (one would expect others), and, second, it wouldn't seem like a very good hypothesis to adopt. The hypotheses that other people can control your actions or insert thoughts into your head, or talk to you in their physical absence, ought to be assigned a pretty low probability compared to, say, the explanatory hypothesis that something is wrong with your nervous system.

What if there is a more straightforward story to be told about how the move is made from 'not me' to 'someone else'? The PPF may have the resources to do just that. From birth we learn about the world as our nervous systems become sensitive to statistical regularities, and this is manifested in hierarchically arranged hypotheses and expectations. Two very different kinds of stimuli, about which a (even moderately) developed human being's nervous system will have a host

of different kinds of expectations, are inanimate objects, and animate objects (namely, agents). The expectations our nervous systems have about inanimate objects embody our *naïve physics*, so to speak, the expectation about agents, our *naïve psychology* (e.g. Spelke, 2000). Now, if a phenomenon exhibits basic statistical characteristics that activate our naïve psychology, this will be experienced as the work of an agent, and in a way that the subject may find very hard to override.

This may account for why agency, in a generic sense, is attributed, but as for why it is often *specific* agents, the answer may be as follows. Our nervous systems have expectations about types of thing. It, however, also makes sense that it should have expectations about—representations of—specific individuals (including oneself). Some of these individuals may be particularly salient as a result of the subject's past, or may be constructed and attached to a particular statistical pattern that is recurrently present in the experience. This idea amounts to a sort of merging of theories of agent tracking (see Bullot, 2009) and predictive processing. We see no reason to think that these two theories aren't compatible; indeed, the agent representation that is used in tracking could be viewed within the PPF as a generative model for that specific individual.[6]

Such agent-specific generative models won't only have utility in interacting with others (verbally, visually, etc.) but also in live interactions with oneself, where a generative model of oneself will be active and liable to being updated or diverged from. This would occur in inner speech, among other contexts (and may contribute to fleshing out just what is meant by a misattribution of inner speech, and the different ways in which it can be incurred). Furthermore, such generative models needn't be restricted to live interactions, but would come into play in simulated interactions with others (and indeed oneself). This would also encompass cases of dialogic inner speech where other individuals are represented (see McCarthy-Jones and Fernyhough, 2011).

4.3. Soundless voices

One important feature of the PPF is that hypotheses are hierarchically arranged in terms of how concrete and fine-grained they are. Thus, when we perceive things visually or auditorily, our brains are adopting

---

6 Future avenues for research could tie the disruptions of these agent-specific generative models to delusional misidentification, where people claim that the misidentified person looks the same, but is somehow different. This ineffable difference may be due to changes to expectations about that person, which the person is no longer fulfilling (and hence there is a generation of prediction error).

hypotheses about specific colours and sounds, as well as higher-level hypotheses about, say, tables and chairs (in the visual case) and, say, melodies (in the auditory case). It is plausible to think that there are special intermediary hypotheses involved in linguistic cognition that correspond to specialized areas of linguistic expertise, from phonology, lexicon, grammar, all the way up to the literal and intended meanings of whole utterances. In perception, the low-level hypotheses tend to ground the higher-level ones: you experience a particular sentence because you experience particular words, which in turn you experienced because you experienced particular phonemes, and particular phonemes because particular sounds. Of course, the extraction of these is driven by top-down expectations, as the PPF would suggest. However, the high-level hypotheses tend not to be active in the absence of lower level ones: you don't auditorily perceive words in the absence of perceiving sounds. Things are somewhat different in imagination and in 'thought'. The higher-level hypotheses are activated with degraded or absent sensory hypotheses. That's arguably what more or less 'abstract thinking' is.

But what if something has the externality of a perceptual experience, but has the informational quality of one of these more 'abstract' episodes? That is precisely what we seem to get in the (not especially rare) cases of 'hearing soundless voices'. Higher-level hypotheses are activated, with an unusual perception-like vivacity, in the absence of lower-level ones. This would yield an unfamiliar sort of perception-like experience, in the absence of sensory qualities. Here is a self-report of such an experience:

> It's hard to describe how I could 'hear' a voice that wasn't auditory; but the words the voices used and the emotions they contained (hatred and disgust) were completely clear, distinct, and unmistakable, maybe even more so than if I had heard them aurally. (Woods *et al.*, 2015, p. 326)

This idea of higher-level hypotheses being active in the absence of those lower level ones that usually accompany them is in keeping with work examining the idea that the experience of communication may be at the heart of AVHs. In particular the idea is that sometimes what is experienced is the communicative intention, e.g. the intention to insult, which may or may not bring about an accompanying sensory auditory phenomenology (Deamer and Wilkinson, 2014). In principle, the PPF allows for the separation of the levels of the hierarchy, since the precision can be turned down at any point in the hierarchy, leading to one level no longer being answerable to the other (which is the same principle as decoupling from the world, but occurs within the nervous system itself).

5. Recap and Conclusion

What have we learnt about inner speech? Well, what inner speech fundamentally *is,* when viewed through the lens of the PPF, is the generation by my brain of a decoupled hypothesis that I am speaking (which I am doing for my own cognitive benefit). When I am speaking out loud, there are motoric and proprioceptive elements, and there are also auditory elements. Similarly, when I am engaged in inner speech there is both auditory and motoric imagery (predictions which are united under the same hypothesis—namely, that I am speaking to myself, or at least for my own benefit). How accurate is the hypothesis? Well, the hypothesis is *multi-layered*: there are low-level, decoupled predictions about auditory and proprioceptive stimulation, which, in a sense, are inaccurate, but unproblematically so, since they are deliberately decoupled. They are cases of imagery, not perception. There are also high-level predictions about my own agency and communicative intentions, and *these* are in an important sense *not* decoupled. But in a similar vein, they are also, at least usually, *accurate*: I *am* speaking, performing speech acts, when I experience my healthy, ecologically valid, inner speech. This combination, within a unified hypothesis, of coupled and decoupled predictions, this hybrid of imagination and self-perception, means that inner speech involves a delicate balance. The high-level hypothesis that this is *me*, and that I am saying *this*, is liable to be discarded in favour of another hypothesis (this is someone else, and they are saying something else), if there are disruptions to either aspects of the lower-level sensory and proprioceptive decoupled predictions, or to aspects of more high-level predictions. In particular, these predictions may remain *de facto* decoupled, but not recognized as such by the experiencing subject. This involves a perceptualization of imagery: a percept with perception-like vivacity, but which isn't answerable to what's happening in the world.

References

Adams, R., Shipp, S. & Friston, K. (2013) Predictions not commands: Active inference in the motor system, *Brain Structure and Function*, **218**, pp. 611–643.

Alderson-Day, B. & Fernyhough, C. (2015) Inner speech: Development, cognitive functions, phenomenology, and neurobiology, *Psychological Bulletin*, **141** (5), pp. 931–965.

Badcock, J.C., Waters, F.A.V., Maybery, M.T. & Michie, P.T. (2005) Auditory hallucinations: Failure to inhibit irrelevant memories, *Cognitive Neuropsychiatry*, **10** (2), pp. 125–136.

Bullot N. (2009) Toward a theory of the empirical tracking of individuals: Cognitive flexibility and the functions of attention in integrated tracking, *Philosophical Psychology*, **22** (3), pp. 353–387.

Cho, R. & Wu, W. (2013) Mechanisms of auditory verbal hallucination in schizophrenia, *Frontiers in Schizophrenia*, **4**, pp. 1–8.

Dodgson, G. & Gordon, S. (2009) Avoiding false negatives: Are some auditory hallucinations an evolved design flaw?, *Behavioural and Cognitive Psychotherapy*, **37** (3), pp. 325–334.

Feinberg, I. (1978) Efference copy and corollary discharge: Implications for thinking and its disorders, *Schizophrenia Bulletin*, **4**, pp. 636–640.

Friston, K. (2005) A theory of cortical responses, *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **360** (1456), pp. 815–836.

Frith, C. (1992) *The Cognitive Neuropsychology of Schizophrenia*, Hove: Lawrence Erlbaum.

Frith, C. & Done, D. (1989) Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action, *Psychological Medicine*, **19**, pp. 359–363.

Frith, C., Blakemore, S.-J. & Wolpert, D.M. (2000) Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action, *Brain Research Reviews*, **31** (2–3), pp. 357–363.

Gagnepain, P., Henson, R.N. & Davis, M.H. (2012) Temporal predictive codes for spoken words in human auditory cortex, *Current Biology*, **22** (7), pp. 615–622.

Gould, L.N. (1948) Verbal hallucinations and activation of vocal musculature, *American Journal of Psychiatry*, **105**, pp. 367–372.

Helmholtz, H. von (1866) Concerning the perceptions in general, in *Treatise on Physiological Optics*, Southall, J.P.C. (trans.), 1925, Opt. Soc. Am. Section 26, reprinted, New York: Dover, 1962 (vol. III, 3rd ed.).

Jacobsen, E. (1931) Electrical measurements of neuromuscular states during mental activities, VII: Imagination, recollection, and abstract thinking involving the speech musculature, *American Journal of Physiology*, **97**, pp. 200–209.

Jones, S.R. (2010) Do we need multiple models of auditory verbal hallucinations? Examining the phenomenological fit of cognitive and neurological models, *Schizophrenia Bulletin*, **36** (3), pp. 566–575.

Jones, S.R. & Fernyhough, C. (2007) Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations, *Consciousness and Cognition*, **16** (2), pp. 391–399.

McCarthy-Jones, S.R. & Fernyhough, C. (2011) The varieties of inner speech: Links between quality of inner speech and psychopatho-

logical variables in a sample of young adults, *Consciousness and Cognition*, **20** (4), pp. 1586–1593.

McGuigan, F. (1966) Covert oral behaviour and auditory hallucinations, *Psychophysiology*, **3**, pp. 73–80.

Moseley, P. & Wilkinson, S. (2014) Inner speech is not so simple: A commentary on Cho and Wu (2013), *Frontiers in Psychiatry*, **5**, art. 42.

Pezzulo, G. (2014) Why do you fear the Bogeyman? An embodied predictive coding model of perceptual inference, *Cognitive, Affective, and Behavioral Neuroscience*, **14** (3), pp. 902–911.

Pickering, M. & Clark, A. (2014) Getting ahead: Forward models and their place in cognitive architecture, *Trends in Cognitive Sciences*, **18** (9), pp. 451–456.

Seal, M.L., Aleman, A. & McGuire, P.K. (2004) Compelling imagery, unanticipated speech and deceptive memory: Neurocognitive models of auditory verbal hallucinations in schizophrenia, *Cognitive Neuropsychiatry*, **9** (1–2), pp. 43–72.

Smailes, D., Alderson-Day, B., Fernyhough, C., McCarthy-Jones, S. & Dodgson, G. (2015) Tailoring cognitive behavioural therapy to subtypes of voice-hearing, *Frontiers in Psychology: Psychopathology*, **6**, art. 1933.

Spelke, E. (2000) Core knowledge, *American Psychologist*, **55**, pp. 1233–1243.

Wilkinson, S. (2014) Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework, *Consciousness and Cognition*, **30**, pp. 142–155.

Wilkinson, S. & Bell, V. (2016) the representation of agents in auditory verbal hallucinations, *Mind & Language*, **31** (1), pp. 104–126.

Woods, A., Jones, N., Alderson-Day, B., Callard, F. & Fernyhough, C. (2015) Experiences of hearing voices: Analysis of a novel phenomenological survey, *The Lancet Psychiatry*, **2** (4), pp. 323–331.