# 1

# The Causes and Contents of Inner Speech

*Peter Carruthers*

This chapter will first sketch an account of how inner speech is generated. It will suggest that most inner speech comprises attended "sensory forward models" of mentally rehearsed speech actions. The chapter will then argue that inner speech needs to be interpreted by normal language-comprehension mechanisms in order to acquire content. The contents of inner speech, it will be suggested, can include semantic and pragmatic information ("what is said" and "what is meant"), as well as mental state information (what attitude one takes to the saying of it—judging, believing, wondering whether, and so on).

## 1.1 Causes

Inner speech may not be a unitary phenomenon. On the contrary, we can distinguish two distinct forms of it. Both are, in a sense, active; but only one of the two involves mental rehearsal of action. I begin with the kind that doesn't.

### *1.1.1 Auditory imagination*

Sometimes the voices that we hear in our heads arise from episodic memory (even when those voices are our own). In episodic remembering we can *re-experience* an earlier event (Tulving, 1983; Michaelian, 2016). The component experiences can be visual, tactile, proprioceptive, olfactory, gustatory, auditory, or any combination thereof, together with abstract conceptual information. One might recall an argument with one's spouse, for example, in which facial expressions, gestures, spoken words, and tones of voice all figure. In the course of activating these memories one will both seem to "hear" one's spouse speak, and will also seem to hear one's own voice. On other occasions one might recall *just* the experience of hearing someone speak. In that case the only content to the memory will be an inner voice (whether one's own or someone else's).

Note that (in contrast with the other form of inner speech to be discussed shortly), there is no reason to think that episodic speech memories must be supported by motor-schema activation. (Motor schemata are representations, of varying degrees of abstractness, that initiate and control bodily movements. They are stored in and activated from motor and premotor cortex. They underlie "rote" knowledge such as one's knowledge of familiar phone numbers or one's times tables, and are generally thought to be quite distinct from semantic and episodic memory (see Jeannerod, 2006).) The question is an empirical one, of course. And it might turn out that whenever an episodic memory of a speech-involving event is formed, motor-schema representations form a necessary component of the memory ensemble and its subsequent retrieval. But this would be surprising. No other components of episodic memory seem to have this sort of privileged status.

There remains another sense in which episodic memory for speech is active, however. This is that episodic memories can be actively sought for. One can *try* to remember the exact way in which a sequence of events unfolded, probing one's memory with cues of various kinds. Of course, one often seems passive in the face of one's memories, too. Memories can be evoked by a spoken phrase or a familiar smell, often seeming to *intrude* into one's consciousness, unbidden. Indeed, many evoked memories can be unwelcome ones, which one then tries to repress. Even in such cases, however, there is arguably an underlying process of selective attention, which is best thought of as a kind of action (Carruthers, 2015a).

Memories of various sorts are continually being evoked by aspects of one's current experience without ever becoming conscious. These memories are assessed for relevance to current goals and values by the bottom-up saliency network (Corbetta & Shulman, 2002; Corbetta et al., 2008), competing for the resources of top-down attention. When attention is switched from its current focus, and the memory becomes conscious, this might be thought of as resulting from an unconscious *decision* to attend. Even seemingly passive memories are, arguably, actively made conscious, whether the results are welcome or not. Since the considerations that give rise to a decision-to-attend are a limited subset of one's overall values and concerns, it shouldn't be surprising that once a memory has become conscious, and is thus made available to a larger set of considerations, one might then take a decision to suppress it.

Episodic memories are not just actively searched for or actively selected, but can serve as raw material for an active construction process. One can put together a memory of what a reindeer looks like with a memory of what the roof of one's house looks like to construct an image of a reindeer on one's roof. In creative visual imagination one actively searches for, activates, and combines components of different memories to generate a visual representation of something not previously experienced. Likewise, I suggest, in auditory imagination. A composer might call up from memory fragments of familiar tunes, for example, combining and then recombining them into new melodies. And in the same way one might be able to imagine one's spouse saying something never previously uttered by activating and combining memories of the component parts of that utterance, which *have* been previously experienced. The result would be a token of inner speech that is actively created but doesn't result from motor-schema activation.

I know of no direct evidence on the question, but my guess is that inner speech that results from auditory constructive imagination is comparatively rare. This is because most ways of combining together remembered fragments of speech will result, not in speech, but in nonsense. (There are far fewer constraints on the ways in which one can spatially recombine visual images or temporally recombine musical notes and phrases.) The only way to guarantee that the results of such combination will be intelligible is to engage in the speech production process oneself, thereby activating the constraints in question. But if one does that, then one is effectively engaged in mental rehearsal of speech actions, of the kind we focus on next.

### *1.1.2 Mental rehearsal*

It is likely that most episodes of inner speech result from mental rehearsals of the corresponding speech actions. They are thus active in a more direct sense than are forms of inner speech that are episodic memories or employ constructive auditory imagination. This is because they implicate the resources of the language-production system rather than just memory and attentional systems. In such cases one formulates the motor plans for a specific speech action, such as uttering the words, "It is nearly time to go home." These motor schemata are used to create a representation of what it would sound like if they were carried through to completion. (More on this in a moment.) But at the same time, the normal effects of those motor instructions on one's articulatory muscles are suppressed. The result is an experience somewhat like hearing oneself say the words in question although one says nothing aloud.

How and why do motor plans for speech give rise to representations of the sound of speech? *That* they do so has been known for a long time. Early models of working memory incorporated a "phonological loop," for example (Baddeley, 1986), which is what enables one to keep a phone number actively in mind while searching for somewhere to write it down. And it has long been known that articulatory engagement (such as repeating an unrelated syllable aloud while undertaking the task) destroys performance that relies on the phonological loop. It seems that one keeps the phone number active by initiating, and repeating, the sequence of commands needed to say it aloud. When those commands are displaced by others, the sound of the phone number is largely lost. (At that point one has to fall back on unaided memory.)

There is also broad agreement about *how* motor planning gives rise to inner speech. Motor schemata are selected and activated (more on the selection process in Section 1.1.3), issuing in an "efference copy" or "corollary discharge" of those instructions while downstream activation of the facial and throat musculature are suppressed. The efference copy is used to create a sensory representation of what the speech act in question would have sounded like had it been carried through to completion (Tian & Poeppel, 2010; Corley et al., 2011; Scott, 2013; Scott et al., 2013). Indeed, we know from recent work using brain-scanning of various kinds that the process proceeds in two stages, with motor activations first creating a somatosensory representation of what the movements in question would feel like, which then creates an auditory representation of what the resulting speech would sound like (Tian & Poeppel, 2013). The process begins with activity in Broca's area (the classic speech production/speech-planning area), which then activates the relevant regions of primary motor cortex. This then causes activity in somatosensory regions of parietal cortex, before activity is produced in abstract auditory regions of the superior temporal gyrus and superior temporal sulcus. (Note that the latter are also speech comprehension areas. More will be said about the role of speech comprehension during production in Section 1.1.3.)

It might be objected that the work of Oppenheim & Dell (2008, 2010) counts against this picture. The authors show that the sorts of errors that emerge when tongue-twisters are rehearsed in inner speech suggest that only non-articulatory elements of speech are normally implicated. They argue that inner speech normally only comprises abstract phonological information (unless, that is, participants are asked to "silently mouth" the tongue-twisters in question). But this objection results from a misunderstanding of the findings. For phonological-planning is still motoric, albeit at a more abstract level than the sort of motor-planning that would include phonetic details of accent, stress, and so forth. And indeed, it is well known from the action-planning literature generally that planning always involves a cascade of activation from more abstract to more concrete levels, with efference copies being created at multiple levels. (See the discussion in the paragraph that follows.) Moreover, in their review of the literature on inner speech, Dell & Oppenheim (2015) make clear that they think inner speech results from rehearsed motor plans, which are generally fairly abstract, but can also be quite concrete, depending on the circumstances. (For example, an actor using inner speech to rehearse for her part in a play will likely include phonetic detail as well as more abstract phonology.) And it is worth noting, too, that a similar distinction between levels of representation can be drawn for other forms of sensory imagery also. Thus, one can imagine a reindeer on one's roof with, or without, including details of color and surface textures in addition to shapes and spatial layout.

The question of *why* it is that motor planning can be used to give rise to representations of the sound of speech is somewhat harder. Many have thought that this might be an exaptation of mechanisms initially put in place for purposes of motor control generally (Blakemore et al., 2000; Frith et al., 2000; Carruthers, 2011). Theoretical models of skilled action suggest that efference copies of motor instructions are used to create sensory *forward models* of the likely consequences of those instructions, at multiple levels of representation. These are then received by a comparator mechanism where they can be lined up against the actual afferent feedback as the action unfolds, on the one hand, as well as against the goals of the action, on the other; thereby enabling swift online corrections in cases of mismatch (Wolpert & Kawato, 1998; Wolpert & Ghahramani, 2000; Jeannerod, 2006). For example, when reaching to pick up a water jug, expecting it to be full, one will have generated strength-of-grip and upwards-force commands appropriate for that expectation. If the jug turns out to be empty, however, the afferent feedback will fail to match the sensory forward model, and one will adjust one's grip strength and upwards force accordingly (and often unconsciously).

Sensory forward models are thought to provide the basis of so-called motor imagery (Jeannerod, 2006). (Strictly speaking, motor imagery is proprioceptive and kinesthetic imagery, since the forward model is a prediction of the sensory consequences of movement, which are made available via these senses.) One can activate the motor instructions for a dance move, for example, or a ski turn, while suppressing the downstream movement itself. The efference copy of those instructions is used to create a sensory forward model of the afferent feedback that should occur if the instructions were executed, which will be in proprioceptive and kinesthetic codes; and these, when attended, become available to consciousness. But sensory forward models are not restricted to somatosensory modalities. Efference copies can also be used to create visual representations of the trajectory one's hand should take toward the jug, and the speed with which it should approach it; again allowing for swift online correction using visual feedback. Inner speech, on this account, is just a sensory forward model in auditory code produced by activated (but not executed) speech actions.

While sensory forward models are a necessary condition for the occurrence of inner speech, they aren't sufficient. For their primary use in motor control requires that they co-occur with actual movement and actual afferent sensory feedback. For inner speech to occur, there needs to be a capacity to suppress motor activity downstream of efference-copy creation, and this needs to be coordinated with attention directed at the forward model. These requirements are another matter entirely, and seemingly play no role in motor control. Their use, rather, seems to be in reflective (conscious) forms of motor *selection*. By activating the motor instructions for a potential action, while repressing its execution, one can evaluate the likely further consequences and likely success of the action. The sensory forward model, when attended and globally broadcast to the full range of inferential and affective-evaluative systems, can be elaborated and evaluated in advance of (often instead of) its execution. By mentally rehearsing potential actions, one can evaluate them and make selections among competing plans. Indeed, this is often what one does when engaging in so-called "prospection" of the future (Gilbert & Wilson, 2007).

There is evidence that capacities for mental rehearsal of action are widespread in the animal kingdom, and are commonly found among mammals and birds at least (Carruthers, 2013, 2015a). Let me describe just one experiment conducted with New Caledonian Crows by Taylor et al. (2010), which provides a particularly spectacular example. The birds were presented with a problem to solve: there was some meat within a clear Perspex container, but so deep in the container that it could only be reached by using a long stick, which was behind the bars of a nearby cage. The long stick was too far behind the bars for the birds to reach with their beaks, but there was a short stick hanging on a string nearby. The birds had previously had experience of using a long stick to extract food from the Perspex container, and also with meat hanging from a string, but not with a stick hanging on a string or with the other components of the task. In order to succeed, the bird would first need to pull up the string to obtain the short stick, then use the short stick to extract the long stick from behind the bars, before using the latter to obtain the meat.

Two of the birds solved the task on the first trial (and all four solved it by the second trial). One of these birds inspected the setup for 110 seconds before acting, and then seamlessly executed the entire sequence. The other inspected the apparatus for forty-three seconds, then started to pull up the string before dropping it again. He then inspected the setup for a further forty seconds before solving the problem. What were the birds doing during the inspection periods that enabled them to construct a three-step plan? It is hard to resist the conclusion that they were mentally rehearsing the actions open to them, and combinations of such actions, until they hit upon the correct sequence. Moreover, we know that birds have the attentional networks needed for sensory representations of the various actions to be globally broadcast and evaluated (Winkowski & Knudsen, 2007, 2008; Lai et al., 2011; Mysore & Knudsen, 2013), and for sure they would possess the networks required for sensory forward modelling of action, since even dragonflies have these (Mischiati et al., 2014).

While many animals are capable of mental rehearsals that would transform motor plans into visual and perhaps proprioceptive format, there is no evidence that other animals use mental rehearsals for planning vocal actions, which require motor-to-auditory transformations. Indeed, it seems that other primates don't have their vocalizations under full voluntary control. Instead, their vocalizations appear to result more directly from affective processes, in the same way that emotionally caused postures and facial expressions do—that is, in ways that can be voluntarily modified and inhibited, but aren't voluntarily produced (Cheney & Seyfarth, 2007).

When humans first began to bring vocalization under voluntary control, then, for purposes of speech, there would need to have been some further selection pressure in place before mental rehearsal of speech actions could become possible. This would likely have resulted in copying across into the motor-to-auditory domain mechanisms for coordinating motor-schema activation, attention, and motor suppression (which already existed in other domains). (Duplication of mechanisms of this sort is quite common in biology, and is one of the main engines of evolutionary change; see Kaas, 1989; Zhang, 2003; Barrett, 2012.) Plausibly, the adaptive pressure, here, would have derived from the benefits of planning, just as it does with other forms of mental rehearsal. Since speech, from the outset, would have been of vital social (and hence adaptive) significance, one would expect that people might benefit from rehearsing some of their speech actions in advance, to gage their likely social effects.

I tentatively suggest, then, that the evolutionary function of inner speech is to enable the mental rehearsal and evaluation of overt speech actions. Using inner speech, we can try out saying things in advance of saying them, thereby estimating and evaluating their likely effects. And for sure, people sometimes do use inner speech in this way. Someone considering a marriage proposal, for example, might rehearse various ways of "popping the question" to select the one most likely to be effective. Likewise, many of us rehearse and evaluate potential ways of expressing our ideas when planning a lecture. Of course, not all inner speech nowadays is directed at an imaginary audience, and no doubt it is used for many purposes in addition to its original planning function. Indeed, it may well have become co-opted in normal development (either by having come under secondary natural selection, or through cultural evolution) to play the sorts of self-management roles postulated by Vygotsky (1961) and others. But if one asks what inner speech is *for* (in the sense of initially being *adapted for*), it seems likely that the answer is: overt, conscious speech planning.

I have argued that most inner speech is actively produced. But how can this account accommodate the robust finding from introspection-sampling studies that people frequently report a distinction between "inner speaking," which is active, and "inner hearing," which isn't (Hurlburt et al., 2013)? One possibility is that all instances of inner hearing result from forms of episodic auditory memory, of the sort discussed in Section 1.1.1. But this seems unlikely. More importantly, the fact that people sometimes don't *experience* their inner speech as active simply has no bearing on the question whether it is actually produced by motor-schema activation. For the latter is always unconscious. It may be that the distinction reflects, rather, whether or not there is some degree of awareness of somatosensory activation in inner speech, taking such awareness as a cue to self-produced activity. The experienced distinction between inner speaking and inner hearing may then result from subtle differences in the direction of attention (whether toward auditory sensory areas alone, or also to the somatosensory forward models implicated in production), that is all.

### *1.1.3 Inner speech selection*

It seems that most inner speech results from rehearsed motor plans. But what selects *which* motor plans are rehearsed? A reasonable assumption is that the selection process overlaps, at least, with those involved in overt speech production. And here the traditional view has been that speech begins with a *message to be communicated* (Levelt, 1989). This would be a propositional content of some sort (a thought) that gets motorically encoded, first by selecting lexical items for an appropriate syntactic frame, and then using that to generate phonological plans and lower level motor instructions. If so, then inner speech begins with a thought that is already fully formulated. But this view is arguably too simple (even if it was once a useful idealization). In fact, speech production in general (like speech comprehension; Hickok & Poeppel, 2007) seems to proceed in parallel (or at least interactively; Nozari et al., 2011), with decisions about *what* to say being taken while one is in the process of saying it (Dennett, 1991; Lind et al., 2014). If so, then one generally doesn't *first* formulate a determinate message *before* expressing it in speech; rather one often formulates messages *by* expressing them into speech (that is to say, in the course of doing so), perhaps involving competition among multiple possibilities.

One consideration supporting this proposal is provided by extant models of motor control quite generally. For forward-modelling accounts always make provision, not just for comparisons between sensory forward models and afferent feedback, but also for comparisons between forward models and the goals of the action. This makes it possible to make motor adjustments in advance of afferent feedback in cases where the movements selected are in some way inadequate or suboptimal. But they also enable one to make adjustments in lower level goals in the light of higher level ones. In the case of complex actions that unfold over time (such as the production of a spoken sentence), one might expect that detailed goals (which words to select, in which order, with what tone of voice) would get selected serially, involving continual interactions between one's goals at the most abstract level and the forward models produced as those goals are progressively implemented. (Consistent with this suggestion, when people participating in introspection-sampling studies are "beeped" in the middle of saying something, they most commonly report that they don't know how the sentence in question would have finished (see Hurlburt & Schwitzgebel, 2007).)

Moreover, we know from decades of work in social psychology that there are generally multiple motivational factors at work, influencing how people express themselves in speech. There are dissonance effects and self-presentation effects, for example, and people may modify what they say when they anticipate disagreement from their audience. (See Carruthers, 2015a, for discussion.) Note that most of this literature concerns how people choose to answer a single determinate question, such as, "How opposed would you be to a rise in tuition costs next semester?" We can assume that in free conversation there will be many more factors influencing what people choose to say in the moment (including a variety of communicative goals, memories or feelings evoked by the circumstances, as well as factors in the immediately preceding conversation).

These suggestions are consistent with the data reported by Novick et al. (2010) that patients with damage to Broca's area (leading to a form of production aphasia) also show much wider deficits, especially in their capacity to inhibit prepotent actions. (For example, they perform quite poorly in the Stroop test.) For the expressive difficulties experienced by some of these people emerge most clearly in cases where there are many competing things they could say. For example, when asked to generate verbs associated with a given noun, patients with damage to Broca's area may become paralyzed when prompted with "ball," since there are many related verbs to choose from ("throw," "kick," "bounce," "pass," "catch," and so on). But they may perform much better when prompted with "scissors," which is associated with just a single action ("cut"). Similarly, healthy people given the same test show increased activity in Broca's area when selecting a verb out of many alternatives, as well as during conflicting-action trials of the Stroop test. At the very least these findings establish that speech production involves competition among expressive *actions*, if not competition among intended messages.

The latter claim is supported, however, by findings from patients with Wernicke's aphasia (who often have severe speech comprehension difficulties), as Langland-Hassan (2015) points out. For although the speech of such patients can be fluent, it is often garbled or completely unintelligible, containing misused words, non-words, and meaningless concatenations of the "Green ideas sleep furiously" variety (LaPointe, 2005). Since Wernicke's aphasia is primarily a speech comprehension deficit, we can infer that speech production normally proceeds in parallel with comprehension, evaluating the contents attaching to a range of potential speech actions while they are being constructed and implemented. (See also Matsumoto et al., 2004; Aristei et al., 2011; Pickering & Garrod, 2013.) In fact, this may be the most common use of the forward-modelling and planning networks discussed in Section 1.1.2, with partially constructed motor plans generating sensory forward models that are semantically evaluated by the comprehension system and assessed against the abstract goals involved in the exchange. Without attention directed at the sensory forward models in question, however, the process will proceed unconsciously (as well as quite swiftly, given the speed with which normal speech can be generated).)

(It is worth noting that forward models might be constructed and compared with abstract goals at multiple levels, and not just at the final—sensory—level. For example, there might be a forward model and comparator process at the stage of lexical selection, enabling one to correct an erroneous choice of word at an early stage in the production process (see Hartsuiker, 2014). Such models are incapable of becoming conscious, however, since consciousness always depends on attention directed at sensory representations of some sort, resulting in their global broadcast. (See Carruthers, 2015a, for extended discussion and defense.)

If we take outer speech production as our guide, then I suggest that inner speech (like outer speech) will be governed by an unconscious decision-making process. A number of goals, salient values, and ways of satisfying those goals in speech will be in competition with one another, influencing the production of sensory forward models that are evaluated unconsciously by the language comprehension system for relevance and appropriateness. When one formulation of any given component of the unfolding speech act wins the competition and gets selected, the others are suppressed, and attention is directed toward the sensory forward model that remains. The latter becomes conscious, and becomes a component of the item of inner speech one is aware of. In fact, all one is ever aware of is the product, not the process. Once it gets made conscious, however, the speech act in question becomes available to a much wider range of predictive and evaluative processes.

Putting together these points with those arising from our discussion of conscious mental rehearsal in Section 1.1.2, we can conclude the following. When one is using inner speech to rehearse potential public utterances, there will be a *two-stage* evaluative process. First, there is local competition among possible ways of implementing the most abstract goals of the speech act in question. This takes place unconsciously, with the sensory forward models from the winning components attracting attention. The latter become conscious and are thus further evaluated by a wider range of inferential and affective systems (including their likely impact on the mental states of the imagined audience), if necessary leading to iterative conscious rehearsal of yet other possibilities.

There is one significant exception to the account of inner speech production just sketched. This concerns speech actions that have been learned by rote. When one has been drilled to repeat the Oath of Allegiance, for example, the result is that the motor instructions for the entire sequence get chained together and stored as a single unit in motor cortex. When one embarks on the process of saying the Oath (whether aloud or in inner speech) the sequence will unfold without further decision-making. This is just as it is with other forms of habitual action. Once learned and stored, when habits become activated by a cue (either by something in the environment or by the previous action in a habitual sequence), the action is selected and performed without the usual decision-making processes (Wood & Rünger, 2015). A habitual act can be *inhibited* by a decision, but isn't initiated by one.

## 1.2 Contents

I have sketched how inner speech is produced. But what sort of *content* does inner speech possess? And how is that content determined? It will prove helpful to discuss these questions in two stages—first as they relate to our understanding of the speech of another person, before turning to the case of inner speech for comparison.

### *1.2.1 Outer speech*

As we noted in Section 1.1.3, speech comprehension seems to take place interactively, across all levels of representation and drawing on background knowledge and expectations, rather than sequentially (Hickok & Poeppel, 2007; Aristei et al., 2011; Sohoglu et al., 2012). The process has to begin with patterns of sound stimulating one's eardrums, of course, giving rise to initial stages of spectral analysis. But thereafter hypotheses about lexical items inform hypotheses about phonology, hypotheses about syntax and semantics inform the selection of lexical items, and hypotheses about speaker intent (partly guided by contextual factors and nonverbal cues) inform the selection of syntax and semantics. Of course, one cannot just hear someone as saying one thing rather than another at will. Interpretation is importantly constrained by bottom-up information. But as the McGurk effect demonstrates, even phoneme perception can be influenced by other factors (in this case, by visual perception of movements of the speaker's mouth).

In this respect, speech perception is no different from other forms of perception. Vision, too, has been claimed by many to be deeply interactive at many different levels, including conceptual ones (Rauss et al., 2011; Clark, 2013; Panichello et al., 2013; Vetter & Newen, 2014). This is especially important where visual input is ambiguous, incomplete, or degraded. Conceptual information is used to "query" processing at lower levels, attempting to find a best match for the stimulus, filling in missing contours or other details as necessary. Consistent with this account, Wyatt et al. (2014) review a range of results suggesting that feedback from inferotemporal conceptual areas of cortex has an impact on processing in visual cortex as early as 100 milliseconds following the onset of a stimulus, and significantly before top-down attention can have had any impact (which starts at around 200 milliseconds). Among the effects of this recurrent processing are a number of types of change at lower levels of visual representation, including the filling in of missing portions of a figure that is partly occluded (or that is assumed at higher levels to be partly occluded).

Ogilvie & Carruthers (2016) draw on a range of recent evidence to provide an extended argument for the conclusion that early visual processing is not strictly encapsulated from higher level concepts and background knowledge. But the conclusion they draw is a modest one, and is consistent with a continuing distinction between the visual system proper (comprising cortical regions specialized for processing signals from the retina) and cognition (which is amodal, or at least multimodal, in nature). To say, as they do, that visual processing and conceptual knowledge interact is not to deny the distinction between them.

Carruthers (2015a) advances an additional argument, again focusing initially on the case of vision. We know that visual processing takes place in a distributed fashion, with color being processed separately from shape, and each being processed independently of movement. Yet each of these separate properties can be bound together into a single percept of, say, a round red object (a tomato) rolling along a surface. (Note that we have known for a good many years that binding can occur prior to and independently of consciousness; Dehaene et al., 2004.) A central organizing principle in the binding process are so-called "object-files" (Pylyshyn, 2003). These are like indexical links to an object ("*That thing*...") to which property information (color, shape, and the rest) can be attached. Carruthers (2015a) argues that the best account of seeing *as* (where the round red object is seen *as* a tomato, for instance) is that category information can be bound into these object-files and then globally broadcast along with them, constituting a single conscious visual percept.

An alternative view is that there are two distinct conscious events: one is a perceptual object-file ("*That* round red rolling thing") whereas the other is a perceptual judgment ("*That* is a tomato"). Notice, however, that such a view faces a new version of the binding problem. For it fails to explain what secures the coincidence of reference of the two indexicals, making it the case that one sees the round red rolling thing as the tomato, rather than something else in the visual field. Moreover, it also predicts that conceptual judgments should be capable of being globally broadcast (thereby becoming conscious) independently of any related conscious percept. Carruthers (2015a) amasses a variety of evidence, and presents a number of arguments, supporting the conclusion that this prediction is false. So the best account of the contents of vision is that visual object-files can contain both conceptual and nonconceptual information.

When we turn to speech perception, the relevant organizing principle is the *event-file* (Hommel, 2004; Zmigrod et al., 2009). (Note that an object-file structure is unlikely to work here, since the only relevant object would be the speaker. But one can understand speech, and bind it into a single interpreted utterance, without knowing or otherwise perceiving the identity of the speaker.) Speech is segmented into distinct events (generally sentences), with multiple properties drawn from many different levels of processing bound into each event-file. Thus, one hears the tone of voice, the volume, and the accent with which someone says something, while also hearing what they say, and often also the intent with which they say it (as when one hears someone as speaking ironically, for example).

What are the constraints on the sorts of information that can be bound into an auditory event-file? These are mostly temporal in character, I suggest. Only information that can be extracted in the first 300–500 milleseconds after stimulus offset, before the experience gets globally broadcast, can become conscious along with the remainder of the file. (On the timing of consciousness, see Dehaene, 2014. Note, however, that since auditory events unfold over time, much of the information can be extracted more slowly before the speech act is completed.) This will be in large part a function of background knowledge and expertise. (Compare how, in the visual domain, a chess grandmaster might be capable of immediately *seeing* White as having a winning position, whereas others can only infer it.) When someone says something unexpected, without context, one might not be able to extract its semantic content. (These are the cases where one responds, "I'm sorry, I didn't catch that" or, "Could you say that again?") In other instances, one might get the semantic—literal—meaning, but not recognize the intent with which the person is speaking. Consider a case where someone says, speaking ironically, "It is a great day for America." One hearer, who knows the speaker well, might immediately hear him as meaning the opposite. Another, with a different set of background assumptions, might hear it literally. And yet another person might be unsure.

Note that, on the account I am proposing, we can often hear people's mental states, just as we can often see them (Carruthers, 2015b). You can hear someone as wanting something, or intending something, or deciding something. And you can—quite literally—hear someone as believing or judging something. (Such experiences are sometimes inaccurate, of course, and always result from contextually sensitive forms of interpretation.) Provided that the mind-reading system can do its work and extract the mental state information from the speech act fast enough, the result can be bound into the event-file and globally broadcast as a component within it.

### *1.2.2 Inner speech: comprehension*

In light of the previous discussion, our main question is whether the content of inner speech depends on the same comprehension mechanisms that issue in the heard content of outer speech. I shall argue that it does, with one type of exception. Let me deal with the exception first.

We noted in Section 1.1.1 that inner speech can sometimes be activated directly from episodic memory. In some of these cases it may be that the intended meaning is activated directly from memory also. This is because episodic memory can link and store information of all kinds. When you recall your twenty-first birthday party, for example, you may call to mind where you were, how the room was decorated, who was there, what you felt, what was said, and so on. Both sensory and conceptual information can be bound into the memory and activated together. An episodic memory of someone saying something, then, might include the content of what they said linked to context, tone of voice, and such like. When the memory is activated, the content of the speech might be right there in the memory itself. The language comprehension system doesn't need to get engaged.

Such cases are likely to be quite rare, however. In part this is because episodic memory for speech is comparatively rare as a proportion of inner speech episodes in general. But it will be even rarer for an episodic memory of someone's speech to be *complete*. Often, one's memory for the original event will need to be reconstructed from more-or-less complete fragments, and it seems inevitable that the language comprehension system would participate in the reconstruction process, selecting from among likely candidates for missing words, for example, and helping to reconstruct the speaker's intent from recalled contextual cues. (On the constructive nature of memory generally, see Schacter et al., 1998.)

What of the much more numerous cases where episodes of inner speech result from mental rehearsal of the corresponding speech actions, then? Here the heard content of the speech can only arise from speech comprehension processes. This is because the efference copy of the motor instructions, which is used to construct the sensory forward model, is just that: a copy of *motor* instructions. It has no intrinsic semantic content. That content will have been left behind in the speech production process that issued in those instructions. The sensory forward model, then, will likewise be just that: a *sensory* forward model. (Recall that even if there are non-sensory—e.g. lexical—forward models involved in the production process, these are incapable of becoming conscious on their own.) The sensory forward model will need to be processed and interpreted for content much as does perception of someone else speaking. Hence, essentially the same multilayer interactive interpretive process will take place here as was outlined in Section 1.2.1 for speech comprehension generally.

Does this conclusion alter at all when we take account of the back-and-forth speech selection process? Recall from Section 1.1.3 that multiple competing choices about ways of implementing a speech act (or the components of a speech act) will generally be in play at any one time, each of which creates its own (generally partial) sensory forward model. The latter are interpreted and evaluated against the background goals and underlying values that initiated the speech act, using speech comprehension processes together with some other predictive mechanisms. By the time a final selection is made, the winning speech act will *already* have been interpreted, in part or in full, in the course of the competitive selection process. But this changes nothing. In selecting each component of a speech act over its competition, the former is fully activated and rehearsed while the remainder are suppressed, and attention is directed at the sensory forward model that results. The content of the latter still depends upon interpretive processes, albeit interpretive processes that figured in the selection of that very content. The content of what one hears, in inner speech, will have been filtered through the same interpretive mechanisms that underlie speech comprehension generally.

Moreover, the back-and-forth evaluation of potential speech act components conducted as the action unfolds is likely to be quite local, probably not including the contribution of the mind-reading faculty to figure out what mental state would be expressed. (Nor is it likely to include unobvious consequences of the speech act in question. Remember, the comparison process is mostly about matching the more abstract goals and values behind the speech act's production against interpreted forward models to guide the production process as it unfolds.) For one doesn't, in general, engage in inner speech in order to learn of one's own mental states. So that won't be one of the dimensions of assessment. As a result, once the content of the selected act is targeted for global broadcasting, mental state attribution may take place for the first time (if at all). (Recall that somewhere between 100 and 300 milliseconds elapses between the impact of attention and a state subsequently becoming conscious. For in the perceptual domain, at least, attention starts to have effects around 200 milliseconds after stimulus offset, whereas global broadcasting takes place around 300–500 milliseconds after stimulus offset. So there is significant time for additional content to be bound into an attended sensory forward model.)

### *1.2.3 Inner speech: content*

Since the content of inner speech results from the same sorts of processes that interpret outer speech, one might expect, on theoretical grounds, that the same range of experienced contents would attach to each. And it seems, introspectively at least, that this is the case, as we will discuss in a moment. One notable difference, however, is that we rarely, if ever, *fail* to attach a content to our own inner speech, in the way that we often fail to comprehend the speech of another. At any rate, I don't think I personally have ever had the experience of asking myself, "What did I just think?" in the way that one often needs to ask others, "What did you just say?" Why this should be so, given that the same comprehension processes underlie each, is an important question, which will be discussed in Section 1.2.4.

I suggest that there are two main categories of inner speech contents. The first is where one hears oneself as expressing a specific semantic content, but without hearing oneself as engaged in any further mental activity. The second is where one hears oneself, in addition, as judging something, deciding something, or whatever. In English these two classes of content would normally be reported to others using distinct forms of "thinks"-sentence. Consider a case where one tokens in inner speech the sentence, "It is time to go home." One might hear this only as what Cassam (2014) calls a "passing thought"—that is, as a content that isn't the object of any particular mental attitude. In such a case one might report the episode by saying, "I was thinking *about* whether it is time to go home." (Note that this isn't the same as saying, "I was *asking myself* whether it is time to go home." Nor is it the same as saying, "I was *wondering* whether it is time to go home." These attribute a particular mental attitude, that of asking a question, or of wanting to know something.) On the other hand, one might hear oneself as *judging* or *deciding* that it is time to go home. Here one would report the episode by saying, "I was thinking *that* it is time to go home."

The factors that determine whether one merely experiences a passing thought or hears oneself as entertaining some specific mental attitude are likely to parallel those that are at work in interpreting the speech of other people. Often there will only be enough information in the utterance and its circumstances to extract a semantic content, not an attitude toward that content. For example, in the public case one might hear the literal meaning of, "It is a great day for America" without knowing whether the statement is ironic, a disguised question, or what. This is like a case of a "passing thought." (An important difference, however, is that we take it there is no such thing as a "passing utterance," for we know that any utterance will express some mental attitude. Why we don't make the same assumption for inner speech is a question we will take up in Section 1.2.5.) But in other circumstances there will be enough surrounding cues, and ones that can be processed swiftly enough, for a mental state attribution to be bound into the content of what one hears. These might include aspects of the content, the speaker's tone of voice, facial expression, and so on. These same cues, as well as others (such as any visual imagery that might co-occur with the utterance), will be available in the case of inner speech also.

### *1.2.4 Why so reliable?*

In light of the immediately preceding discussion, the question that forms the title of this section can be broken down into two separate parts. On the one hand one might ask why we are so reliable in attaching a semantic content of some sort to our inner speech. And on the other hand one might ask why we are so good at discerning (and hearing ourselves as expressing) our underlying mental states. These two questions need to be handled rather differently. Let me start with the one about semantic content.

Why is it that we never seem to fail to attach a content to our own inner speech? We are often at a loss to know what others have just said. But it seems we rarely, if ever, have similar doubts about our passing thoughts. This is true despite the fact that inner speech can often be highly fragmentary, consisting of just a word or a phrase. And in many of these cases, it seems that the propositional content we hear ourselves express isn't obviously given by the surrounding semantic context. (Contrast a case where one hears someone respond to the question, "Will it rain today?" with a simple, "Yes." Here there is no difficulty explaining how one hears the latter as expressing the content, "It will rain today." No doubt similar phenomena can occur in inner speech also. But many cases of fragmentary inner speech with determinate heard content don't seem to fit this sort of profile.)

To answer this question about the ubiquity of semantic content in inner speech we can draw on theories of speech interpretation that place heavy reliance on the *accessibility* of relevant representations (e.g. Sperber & Wilson, 1995, 2002). People will generally understand a pronoun to refer to the most recently mentioned individual, for example, or to the individual who is most salient in the context. And in a conversation with a colleague about where to go in the surrounding city during a lunch break, the sentence, "Can we first go to the bank?" will be heard as referring to a financial institution rather than an edge of a river, because the former interpretation is more easily accessible in the context.

Now put together this point about the role of accessibility in speech interpretation generally, with the account of inner speech production given in Section 1.1.3. Whether one assumes (with traditional models) that production begins from a complete message-to-be-expressed, or rather from a back-and-forth competitive construction process involving speech comprehension systems, the conclusion is the same. The concepts, lexical items, and syntactic structures that are actually deployed in producing a sentence of inner speech will have been maximally active just milliseconds before the interpretive process begins. They will thus be maximally accessible, making it easy for the comprehension system to settle on the correct semantic interpretation. Recall, too, that on the competitive account I prefer, when one component of a speech act is selected over the competing ones, the latter will be *suppressed*. This will mean that the lexical items and syntactic constructions that would have been used in nearby utterances are rendered comparatively *in*accessible.

It might be objected against this proposed explanation that relevance theory is a broadly Gricean account of communication, which thus assumes that successful communication depends richly on communicative intentions (Knappik, 2017). Speakers are assumed to have the intention of producing messages that will be maximally relevant to their audience; and audiences take intentions-to-be-relevant for granted. But in inner speech there generally *is* no communicative intent. Even in cases where inner speech is used to rehearse and evaluate a potential public utterance before making it, there is no actual intention to be relevant.

This objection can be replied to, however. For the intention to be relevant is mostly merely tacit or procedural. (Indeed, reliance on accessibility is a feature of cognitive processes generally, not just communicative ones; see Kahneman, 2011.) Given a communicative goal of some sort (in cases of outer speech), one will, by default, assemble lexical, conceptual, and syntactic materials that are most easily accessible in the context. This works because what is easily accessible to oneself will likewise generally be easily accessible to one's hearer (and will hence be used in the comprehension process). But there need be no intention to select accessible items; one does so, rather, *merely* because that is easiest. Likewise, on the comprehension end of things, a hearer can select interpretations on the basis of ease of accessibility without attributing any corresponding intention to the speaker. By default, hearers will just reach for the easiest interpretation available. These same tacit selection and interpretation processes can then be at work in inner speech, where there is no communicative intent.

It is easy enough to explain, then, why we should almost always settle swiftly on a semantic interpretation of an episode of inner speech, no matter how fragmentary, and why this interpretation should almost always be correct. But what of the mental states we often hear ourselves express? Are we similarly reliable in this? And if so, can we explain how?

As we noted in Section 1.1.3, the processes that underlie speech production in general are opaque to speakers themselves. These include the competing motives and beliefs that influence the selection of any given speech act. As a result, there will often be a significant mismatch between the interpreted intent behind an utterance and the true intention. We tend to think, for example, that when people make literal and non-ironic assertions, and do so without any intention to deceive, that they are thereby expressing the corresponding belief. Someone who says, "It will rain this afternoon" will often, in context, be heard as expressing the judgment or belief that it will rain that afternoon. And the same will be true of our own inner speech. Hearing a sentence as an assertion (and not just a passing thought), we will be apt to hear ourselves as believing it. But the relationship between assertion and belief is far from what this simple picture assumes. We know that multiple motivational factors are generally in play whenever someone makes an assertion (Carruthers, 2011). And many of these—for example, self-presentational or self-enhancement motives, and motives to reduce dissonance—will operate in the case of inner assertion also.

These considerations suggest that one should be no more reliable in identifying the mental states expressed in inner speech than one is when identifying the mental states of others. In both cases one's access is interpretive, not direct; and in both cases there will be a multitude of causal factors intervening between one's speech performances and one's underlying mental states. On the other hand, one might think that some of the additional cues that are available in the first person but not in the third might increase reliability. In the case of another person, for example, one won't have access to any inner speech that they might have engaged in prior to their public utterance, whereas this will be available in one's own case. Likewise, in connection with other people, one has no access to any accompanying or preceding visual imagery, affective feelings, and so on, all of which can (when conscious) be available to assist the interpretation of one's own inner speech.

It is no easy matter to determine the extent of our reliability in attributing mental states to others; and by the same token, it isn't easy to know whether we are any more reliable in attributing mental states to ourselves. In part this is because any apparent inconsistencies we detect between the states we attribute and subsequent or previous behavior may result from changes of mind, rather than errors of attribution. Nor is it easy to detect such inconsistencies in the first place, since behavior is always a product of more than one mental state (minimally, a belief–desire pair). Moreover, once one has been interpreted, by others and/or oneself, as having a given mental state, one will be apt to constrain one's behavior in the future to be consistent with it. As Frankish (2004) points out, when people hear themselves as expressing a given mental state, they may regard themselves as *committed* to having it, and so will constrain their future behavior to be consistent with that commitment. In which case, even if the initial attribution is erroneous, the person may thereafter act *just as if* it had been accurate.

### *1.2.5 Why no uncertainty?*

I have argued that we are highly reliable interpreters of the semantic contents of our own inner speech. In contrast, we are almost certainly not fully reliable when we interpret our inner speech as expressing a specific mental state, and it is hard to know whether we are any more reliable than are others who listen to our overt speech. But there is a different puzzle that also requires discussion. We are sometimes in doubt about the literal meaning of what someone has said. We often hear others as speaking ambiguously, for example, or are otherwise uncertain of what they mean. Likewise, we often aren't sure what mental state to ascribe on the basis of what they say, since their intention in saying what they do can be obscure to us. In contrast, we never seem to have such doubts about our own inner speech. We never seem to hear our own inner speech as ambiguous, for example. And we either hear an inner utterance as a mere "passing thought" (not expressing any mental attitude), or we hear it as expressing a belief, judgment, or whatever; we never seem to vacillate, wondering which it was. Why is this, if our access to the meaning of our own inner speech is just as interpretive as our access to the speech of others?

Our account of why we should never experience inner speech as ambiguous can parallel the explanation given in Section 1.2.4, for why we should never fail to hear a meaning at all. If one employs an ambiguous sentence in inner speech (such as, "I shall go to the bank"), one of the candidate concepts in particular will have been activated during the production process (say, the one referring to a financial institution). In that case it will be highly accessible during the comprehension process, and will become part of the semantic content of the heard utterance with a high degree of reliability and certainty. One should have no sense of competing possible interpretations, in the way that one sometimes does when listening to the speech of another.

A similar approach might enable us to explain why we never seem to doubt whether our own inner speech is meant ironically or not. For there will generally be interpretive cues readily available in the surrounding context. For example, hearing oneself say, "It is a great day for America" when one has been watching—with accompanying feelings of joy—a program about team USA's successes at the Olympics will lead to one interpretation; hearing oneself make the same statement when one has just been watching—with feelings of horror and disgust—a news report about the presidential election will lead to quite another.

In addition, we are, of course, fully aware that interpreting the intent behind other people's speech is a hazardous and highly fallible business. But most of us lack such awareness in the first person. As Carruthers (2011) argues at length, even if people consciously acknowledge a role for unconscious mental states in determining their behavior, they are nevertheless apt to assume tacitly, in online unreflective tasks, that their own minds are transparent to them. Indeed, Carruthers suggests on reverse-engineering grounds that something like a Cartesian model of one's mind as infallibly self-presenting may be built into the structure and unconscious operations of the mind-reading faculty itself, greatly simplifying its functioning. If so, then it is no wonder that we don't hesitate when an interpretation of the intent behind our own inner speech suggests itself.

The same idea can also be used to answer a question we left hanging in Section 1.2.3: why is it that we interpret some items of inner speech as mere "passing thoughts" whereas we don't ever seem to interpret the speech of other people likewise? The explanation may lie in the proposed assumption of self-transparency of mind. For if one did have transparent knowledge of one's own mental states, then of course one would be aware of the intentions behind any utterance that one tokens in inner speech. Since we often find ourselves *unaware* of such intentions (while being aware of a semantic content of some sort) we interpret this as a "free-floating" content or "passing thought." That enables us to preserve the transparency assumption. There is no such pressure, of course, when interpreting the utterances of others. For we know full well that *other* minds are *not* transparent to us (and this, too, is likely built into the structure of the mind-reading faculty). Hence we can continue to assume (correctly) that there must be some determinate intent—some attitude expressed—behind any utterance we hear.

## 1.3 Conclusion

My focus in this chapter has been on the causes and contents of individual items of inner speech, not the extended sequences of such speech that can occupy one's stream of consciousness. The latter is governed by additional factors that have not been considered here. (See Carruthers, 2015a, for discussion.) I have argued that inner speech is, for the most part, caused by the same mechanisms that produce and monitor the production of outer speech; and that the content of inner speech likewise results from the same comprehension mechanisms that interpret outer speech. Inner speech is (normally) an attended sensory forward model of a rehearsed speech action, where that action has been selected over others by unconscious appraisal and decision-making processes. The content of inner speech can include both semantic and mental state information. The interpretive process that results in the former is quite reliable, because of the central role played by *accessibility* in language comprehension generally. In contrast, the process that leads us to hear inner speech as expressing some specific type of mental state is unlikely to be fully reliable (although the extent of its reliability is quite hard to gage).

## Acknowledgments

I am grateful to Peter Langland-Hassan, Agustìn Vicente, and an anonymous referee for their comments on an earlier draft of this chapter.

## References

Aristei, S., Melinger, A., & Rahman, R. A. (2011). Electrophysiological chronometry of semantic context effects in language production. *Journal of Cognitive Neuroscience*, 23, 1567–86.

Baddeley, A. (1986). *Working Memory*. Oxford University Press.

Barrett, H. C. (2012). A hierarchical model of the evolution of human brain specializations. *Proceedings of the National Academy of Sciences*, 109, 10733–40.

Blakemore, S-J., Smith, J., Steel, R., Johnson, E., & Frith, C. (2000). The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. *Psychological Medicine*, 30, 1131–9.

Carruthers, P. (2011). *The Opacity of Mind*. Oxford University Press.

Carruthers, P. (2013). Evolution of working memory. *Proceedings of the National Academy of Sciences*, 110, 10371–8.

Carruthers, P. (2015a). *The Centered Mind*. Oxford University Press.

Carruthers, P. (2015b). Perceiving mental states. *Consciousness and Cognition*, 36, 498–507.

Cassam, Q. (2014). *Self-Knowledge for Humans*. Oxford University Press.

Cheney, D., & Seyfarth, R. (2007). *Baboon Metaphysics*. University of Chicago Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204.

Corbetta, M., Patel, G., & Shulman, G. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron*, 58, 306–24.

Corbetta, M., & Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3, 201–15.

Corley, M., Brocklehurst, P., & Moat, H. S. (2011). Error biases in inner and overt speech: Evidence from tongue twisters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 162–75.

Dehaene, S. (2014). *Consciousness and the Brain*. Viking Press.

Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J.-B., Le Bihan, D., & Cohen, L. (2004). Letter binding and invariant recognition of masked words. *Psychological Science*, 15, 307–13.

Dell, G., & Oppenheim, G. (2015). Insights for speech production planning from errors in inner speech. In M. Redford (ed.), *The Handbook of Speech Production*, Routledge.

Dennett, D. (1991). *Consciousness Explained*. Penguin Press.

Frankish, K. (2004). *Mind and Supermind*. Cambridge University Press.

Frith, C., Blakemore, S-J., & Wolpert, D. (2000). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews*, 31, 357–63.

Gilbert, D., & Wilson, T. (2007). Prospection: Experiencing the future. *Science*, 317, 1351–4.

Hartsuiker, R. (2014). Monitoring and control of the production system. *The Oxford Handbook of Language Production*. Oxford University Press.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8, 393–402.

Hommel, B. (2004). Event files: feature binding in and across perception and action. *Trends in Cognitive Sciences*, 8, 494–500.

Hurlburt, R., Heavey, C., & Kelsey, J. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition*, 22, 1477–94.

Hurlburt, R., & Schwitzgebel, E. (2007). *Describing Inner Experience?* MIT Press.

Jeannerod, M. (2006). *Motor Cognition*. Oxford University Press.

Kaas, J. (1989). The evolution of complex sensory systems in mammals. *The Journal of Experimental Biology*, 146, 165–76.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus, and Grioux.

Knappik, F. (2017). Bayes and the first person: consciousness of thoughts, inner speech and probabilistic inference. *Synthese*, 1–28.

Lai, D., Brandt, S., Luksch, H., & Wessel, R. (2011). Recurrent antitopographic inhibition mediates competitive stimulus selection in an attention network. *Journal of Neurophysiology*, 105, 793–805.

Langland-Hassan, P. (2015). Hearing a voice as one's own: Two views of inner-speech self-monitoring deficits in schizophrenia. *Review of Philosophy and Psychology*, 7(3), 675–99.

LaPointe, L. (2005). *Aphasia and Related Neurogenic Language Disorders*, 3rd edition. Theime Medical Publishers.

Levelt, W. (1989). *Speaking*. MIT Press.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' acceptance of real-time speech exchange indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science*, 25, 1198–205.

Matsumoto, R., Nair, D., LaPresto, E., Jajm, I., Bingaman, W., Shibasaki, H., & Lüders, H. (2004). Functional connectivity in the human language system: A cortico-cortical evoked potential study. *Brain*, 127, 2316–30.

Michaelian, K. (2016). *Mental Time Travel*. MIT Press.

Mischiati, M., Lin, H.-T., Herold, P., Imler, E., Olberg, R., & Leonardo, A. (2014). Internal models direct dragonfly interception steering. *Nature*, 517, 333–8.

Mysore, S., & Knudsen, E. (2013). A shared inhibitory circuit for both exogenous and endogenous control of stimulus selection. *Nature Neuroscience*, 16, 473–8.

Novick, J., Trueswell, J., & Thompson-Schill, S. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, 4, 906–24.

Nozari, N., Dell, G., & Schwartz, M. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, 63, 1–33.

Ogilvie, R., & Carruthers, P. (2016). Opening up vision: The case against encapsulation. *Review of Philosophy and Psychology*, 7, 721–42.

Oppenheim, G., & Dell, G. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, 106, 528–37.

Oppenheim, G., & Dell, G. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory & Cognition*, 38, 1147–60.

Panichello, M., Cheung, O., & Bar, M. (2013). Predictive feedback and conscious visual experience. *Frontiers in Psychology*, 3, #620.

Pickering, M., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36, 329–47.

Pylyshyn, Z. (2003). *Seeing and Visualizing*. MIT Press.

Rauss, K., Schwartz, S., & Pourtois, G. (2011). Top-down effects on early visual processing in humans: A predictive coding framework. *Neuroscience and Biobehavioral Science Reviews*, 35, 1237–53.

Schacter, D., Norman, K., & Koutstaal., W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, 49, 289–318.

Scott, M. (2013). Corollary discharge provides the sensory content of inner speech. *Psychological Science*, 24, 1824–30.

Scott, M., Young, H. H., Gick, B., & Werker, J. (2013). Inner speech captures the perception of external speech. *Journal of the Acoustic Society of America*, 133.

Sohoglu, E., Peelle, J., Carlyon, R., & Davis, M. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, 32, 8443–53.

Sperber, D., & Wilson, D. (1995). *Relevance*. Second Edition. Blackwell.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity, and mindreading. *Mind and Language*, 17, 3–23.

Taylor, A., Elliffe, D., Hunt, G., & Gray, R. (2010). Complex cognition and behavioral innovation in New Caledonian crows. *Proceedings of the Royal Society B: Biological Sciences*, 277, 2637–43.

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, 1, #166.

Tian, X., & Poeppel, D. (2013). The effect of imagination on stimulation: The functional specificity of efference copies in speech processing. *Journal of Cognitive Neuroscience*, 25, 1020–36.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford University Press.

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, 27, 62–75.

Vygotsky, L. (1961). *Thought and Language*. MIT Press.

Winkowski, D., & Knudsen, E. (2007). Top-down control of multimodal sensitivity in the Barn Owl optic tectum. *Journal of Neuroscience*, 27, 13279–91.

Winkowski, D., & Knudsen, E. (2008). Distinct mechanisms for the top-down control of neural gain and sensitivity in the Owl optic tectum. *Neuron*, 60, 698–708.

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3, 1212–17.

Wolpert, D., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11, 1317–29.

Wood, W., & Rünger, D. (2015). Psychology of habit. *Annual Review of Psychology*, 67, 289–314.

Wyatt, D., Jilk, D., & O'Reilly, R. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, 5, #674.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18, 292–8.

Zmigrod, S., Spapé, M., & Hommel, B. (2009). Intermodal event files: integrating features across vision, audition, taction, and action. *Psychological Research*, 73, 674–84.