Towards a Conative Account of Mental Imagery

*This is a preprint. Please cite the published version.*

Philosophers and psychologists assume that mental imagery is a cognitive state, that it represents things as being a certain way. However, I argue that imagery is a conative state: it represents things as to be made a certain way. I challenge the traditional assumption by targeting an increasingly popular cognitive account that identifies mental imagery, such as inner speech, with predictions of sensory input. This predictive account faces both empirical and theoretical problems. The account not only fails to capture the salience effects exhibited by mental imagery, it also cannot explain why subjects should not spend their lives minimizing prediction error by engaging in fantasy. These shortcomings highlight what I call the conative profile of imagery, which includes its salience, motivational, and performance effects. I argue that an alternative, conative view of imagery, according to which mental images are goal states, is best able to account for its conative profile.

Keywords: mental imagery; conation; inner speech; predictive control; predictive processing

## 1. Introduction: Imagery as Cognitive

Philosophers and psychologists assume that mental imagery is cognitive, not conative. Cognitive states represent things as *being* a certain way, while conative states represent things as *to be* *made* a certain way (Shah and Velleman, 2005). For example, beliefs are cognitive states insofar as they aim to conform to the world, while desires are conative states insofar as they aim to have the world conform to them.The assumption that mental imagery is cognitive is in full force when it comes to perceptual images. Visual mental images are taken to represent things as being some visual way; olfactory images represent things as being some olfactory way; auditory images represent things as being some auditory way; and so on. The assumption remains in force in discussions of action. Though some philosophers claim that mental images guide action, it turns out that they do so only in the way that beliefs guide action, by making available information that allows one to successfully execute a desired action (van Leeuwen, 2011). Cognitive psychologists too have stressed that motor images are a window into motor preparation (Jeannerod, 2006). But motor images themselves are treated as cognitive, representing sensory features of an imagined action, “sensations of both effort and movement” (Currie and Ravenscroft, 1997). Philosophers have posited desire-like imaginings (Doggett and Egan, 2007), but these are supposed to have conative features only within the context of pretense. Therefore, like their perceptual kin, action-related images are widely treated as cognitive, not conative.

Against the assumption that mental imagery is cognitive, this paper argues that it is conative. Mental images represent the world as *to be made* a certain way. I motivate the conative account by challenging a specific cognitive account of imagery which holds that mental images are predictionsof sensation. Since predictions represent things as being a certain way, predictive models treat imagery as cognitive. Though my challenge willconcern imagery in general, my focus will be on inner speech, and in particular, auditory verbal imagery. There are a number of interesting questions regarding the relationship between different components of inner speech – propositional, syntactic, phonemic, phonetic, and motor – but for the sake of this paper “inner speech” will refer only to the auditory verbal component (for discussion see Patel (2021)). Inner speech is an apt test case to assess predictive models of imagery because predictive models of inner speech have been more thoroughly developed than predictive models of other kinds of imagery.

In Section 2, I begin by presenting two predictive models of inner speech, the ‘Standard Model’ and the ‘Predictive Processing Model’. I challenge the Standard and Predictive Processing Models of inner speech on both empirical and theoretical grounds. In Section 3, I survey a swath of studies taken to support both models. I argue that not only do these studies fall short of supporting these models, but that both models fail to account for the effects of imagery on salience. Section 4 focuses on the Predictive Processing Model of inner speech. I show that predictive processing models of imagery lead to a problem analogous to the much-discussed ‘dark room’ problem: they fail to explain why subjects should not spend their lives minimizing prediction error by engaging in imagery rather than perception and action. Having criticized predictive models of imagery, in Section 5 I sketch an alternative, goal-based model of imagery on which mental images are conative states. I conclude in Section 6.

## 2. Two Models of Inner Speech

Action involves the execution of motor commands. However, sometimes our limbs err in executing a motor command, as when one fails to grasp the handle of a mug. To correct these errors, there needs to be a mechanism that is able to determine whether an executed limb movement is correct. According to the standard framework of motor control, predictions serve as a standard against which a given limb movement can be checked as correct or incorrect (Wolpert and Flanagan, 2001). On the standard framework, limb movement begins with an intention to move, which is transformed by an ‘inverse model’ into a set of motor commands, which are then sent to the limbs for execution (left-hand side of Figure 1). Prior to the execution of the movement, however, a copy of the original motor command is generated – an ‘efference copy’ – and transformed by a ‘forward model’ into a prediction of the sensory consequences of executing the limb movement (right-hand side of Figure 1). This prediction is compared with the actual sensory feedback generated from executing the movement (bottom of Figure 1). If the actual sensory feedback matches the prediction, the limb movement is considered successful. If there is a mismatch, the difference between the predicted and actual sensory feedback is computed as error and fed back into the inverse model, where motor commands are fine-tuned and the limb movement is re-executed. Predictions thus allow for the detection and subsequent correction of errors in limb movements.

 Figure 1. Standard comparator model of motor control (see, e.g., Wolpert and Flanagan, 2001)

This framework of motor control has inspired an influential family of models of inner speech, which together comprise what I call the ‘Standard Model’ of inner speech. Although there remains disagreement about the specifics, a number theorists have adopted the Model, including Pickering and Garrod (2013), Swiney and Sousa (2014), Vicente and Martinez-Manrique (2016), and, more recently, Carruthers (2018), Lœvenbruck et al. (2018), and Grandchamp et al. (2019). According to the Standard Model, inner speech is identified with the output of a forward model, namely, with a prediction of the sensory consequences of executing speech motor commands (Figure 2). On this view, an intention to produce speech sounds is transformed into a set of speech motor commands, but unlike in normal speech production, the commands are suppressed at the vocal tract (gray). Despite this suppression, an efference copy of the motor commands is still generated, which is transformed by a forward model into a prediction of the speech sounds that would have been produced by the execution of those motor commands. This prediction of speech sounds in the absence of actual speech is experienced as inner speech (green).

Although the Standard Model is easily generalizable to motor imagery – by replacing vocal tract commands with limb movement commands – it has not been applied to other forms of imagery, including visual, olfactory, and auditory imagery. There is, however, an alternative framework that claims to offer a predictive account of imagery in general: the predictive processing framework. Several influential theorists have recently explored the possibility of modeling inner speech using predictive processing (Wilkinson and Fernyhough, 2017). While predictive processes are just one type of process alongside a number of others within the standard motor control framework, the predictive processing framework seeks to understand the mind in terms of predictive processes alone.[[1]](#footnote-1)

 Figure 2. The Standard Model of inner speech (e.g., Pickering and Garrod (2013); Carruthers (2018); and Loevenbruck et al. (2018))

There are three basic components that make up the predictive processing framework: *prediction error minimization*, *prediction error*, and *precision*. First, according to the framework, the mind has the sole aim of *minimizing prediction error*. Though the standard framework also implicates the minimization of prediction error, it incorporates other processes as well. *Prediction error* is the difference between prediction and incoming data. Prediction error is also implicated in the standard framework. However, predictive processing emphasizes a statistical gloss on the notion. Prediction and data are represented by probability distributions, while prediction error is computed as the difference between the means of the distributions. The mind minimizes prediction error by updating either its predictions or data in a way that minimizes the difference between the means of the respective probability distributions.

The third component marks one of the novelties of predictive processing, and will become important in our assessment of its ability to account for imagery. The *precision* of a signal is represented by a probability distribution’s degree of concentration over the mean. In effect, the precision of a signal determines the *weight* or *importance* of that signal in updating prediction or data. If the precision of a prediction is higher than the precision of the data, the update will be more influenced by the prediction than the data, and vice versa if the precision of the data is higher than the precision of the prediction. In short, the precision of a signal represents confidence in a signal, and updates are shaped by that confidence (Hohwy, 2013).

Using these three components alone, the predictive processing framework attempts to model a diverse range of mental states and processes. Perception and action, according to the framework, are simply ways in which prediction error is minimized. Given some prediction-data pair, one way of minimizing prediction error is by revising the prediction to fit the data. This form of prediction error minimization corresponds to perception. In action, we minimize prediction error by changing the data to fit the prediction. One acts so as to bring the world in line with one’s prediction. Action is not caused by independent motor commands, as on the standard motor control framework, but is rather elicited as a means of confirming prior predictions, thereby minimizing prediction error (Hohwy, 2013).

But what determines whether a subject should perceive or act? According to most predictive processing theorists, action is generated when the *precision* of the prediction is sufficiently high relative to that of the incoming sensory data, such that the prediction becomes nearly unrevisable (e.g., Hohwy, 2008). In effect, one is sufficiently confident in the prediction that one does not revise it, and one lacks sufficient confidence in the data that it becomes open to revision. The result is that one is forced to minimize prediction error by revising the data through action. Whether one engages in perception or action is therefore a function of one’s relative confidence in the prediction and data.[[2]](#footnote-2)

Inspired by the predictive processing framework, Wilkinson (2014) and Wilkinson and Fernyhough (2017) explore the possibility of what I call the ‘Predictive Processing Model’ of inner speech. Their account derives from Clark’s predictive processing model of imagery:

The proposal is that the brain, in order to simulate future unfoldings, must mute the weighting on select aspects of the proprioceptive prediction error signal. Suppose this is done while simultaneously entering a high level neural state whose rough-and-ready folk-psychological gloss might be something like “I reach for the cup.” Motor action, on the [predictive processing] account, is entrained by proprioceptive expectations and cannot here ensue. But all the other intertwined elements in the generative model remain poised to act in the usual way. The result should be a “mental simulation” of the reach and hence an appreciation of its most likely consequences. (2013, p.2)

Imagery is achieved, according to Clark, by generating a high-precision prediction – *there are sensations of reaching for the cup*. In the case of action, the high-precision prediction is compared to the data – *there are no sensations of reaching for the cup* – and the prediction error is minimized by reaching for the cup. However, according to Clark, in the case of *simulating* reaching for the cup, the ‘weighting’ on the prediction error is ‘muted’. The prediction error is thereby minimized without updating either the prediction or the data. Because the precision remains high for the prediction that *there are sensations of reaching for a cup* and low for the incoming sensory signal that *there are no sensations of reaching for a cup*, I enjoy motor imagery of reaching for a cup. The crucial feature of Clark’s account of mental imagery is that the precision of prediction error can be modulated, so as to minimize prediction error without updating either the prediction or the data.

Wilkinson and Fernyhough (2017) adopt this style of account for inner speech:

[imagery in inner speech] 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. (p. 296)

Inner speech, according to Wilkinson and Fernyhough, is identical to a high-precision prediction – *there are incoming speech sounds in my own voice*. In speech production, this prediction is compared with the incoming sensory data – *there are no incoming speech sounds in my own voice* – and the subsequent prediction error is minimized via production of those speech sounds. However, in inner speech, the resulting prediction error is minimized via decreasing the precision of the prediction error. Inner speech is identical to the high-precision prediction.

We thus have two predictive models of inner speech – the Standard Model and the Predictive Processing Model. Both claim that inner speech is a prediction of sensory data. Both also claim that inner speech occurs when these predictions are taken offline. More generally, both models view inner speech as a kind of cognitive state, a prediction of speech sounds. However, the models crucially differ in what they say about how predictions are taken offline. The Standard Model claims that predictions are taken offline by suppressing motor commands at the vocal tract. In contrast, the Predictive Processing Model claims that predictions are taken offline using predictive processes alone, namely, by decreasing the precision on prediction error. This difference will become important in Section 4. But first I want to examine whether the empirical predictions made by these two models hold water.

## 3. Empirical Problems with Predictive Models of Inner Speech

Much empirical research has claimed to support predictive models of inner speech. Both models involve a mechanism that determines whether there is a match or mismatch between prediction and sensory data. Such mechanisms enhance processing of the data relative to baseline if the data fail to match the prediction, and attenuate processing if the data matches the prediction. Thus speech feedback is enhanced when its pitch is artificially altered (and so unpredicted) relative to when it is unaltered (and so predicted) (Behroozmand et al., 2011), a finding that has been replicated across a number of experimental paradigms and auditory features (e.g., Heinks-Maldonado, et al., 2005). In order to establish that inner speech is identical to a prediction, then, studies attempt to show either that sensory data that match inner speech are attenuated or that data that fail to match inner speech are enhanced. More precisely, studies have focused on supporting the following prediction:

*Attenuation*: The processing of an incoming sensory signal will be attenuated when it matches inner speech either relative to when there is a mismatch between inner speech and an incoming sensory signal or relative to when there is mere perception of an incoming sensory signal.

*Attenuation* has been investigated using electroencephalography (EEG). EEG is a technique in electrophysiology designed to measure stereotyped electrical potentials produced by the brain. The potential associated with attenuated processing of a stimulus is known as the N100, a potential evoked by an auditory stimulus 100 milliseconds after stimulus onset (Ford, 2001a). I will now present two EEG experiments that claim to confirm *Attenuation* before casting doubt on their support for predictive models of inner speech. I’ll then show that a third, behavioral experiment that claims to validate *Attenuation* does not in fact do so. Although there are a handful of other studies that test *Attenuation*, the experiments that follow have been widely cited and are representative of experimental paradigms used in the studies I am not discussing (Tian and Poeppel (2013), Ylinen et al. (2015), Jack et al. (2019)).

Whitford et al. (2017) asked subjects to sit in front of a ticker tape display and fixate on a line moving across it (Figure 3). The moment the moving fixation point intersected a target, subjects were instructed to produce a speech sound in inner speech for which they were previously cued, e.g., /ba/. At that precise time, a speech sound was also played over their headset. The audible speech sound either matched or did not match the speech sound produced in inner speech. For example, subjects produced /ba/ and heard /ba/ or produced /ba/ and heard /bi/.

 Figure 3. An illustration of the Whitford et al. (2017) ticker-tape paradigm. The cloud indicates inner speech and the diamonds indicate the auditory probes.

Using EEG, the authors found that the speech sound played over the headset was attenuated (relative to baseline listening) only when there was an exact match between the speech sound produced in inner speech and the presented speech sound (i.e., /ba/-/ba/). The authors conclude that the result confirms *Attenuation*: “inner speech is associated with a…content-specific internal forward model” (p. 3).

Whitford et al.’s ticker-tape paradigm is an instance of a paradigm first developed and used by Judith Ford 2001a, 2001b, and Ford and Mathalon (2004). Ford et al. (2001a) tracked attenuation in two conditions. In the baseline listening condition, subjects fixated a target on a screen and were presented probes that alternated between a visually presented checkerboard, speech sounds, and broadband noise. In the inner speech condition, subjects were presented with these same probes while at the same time repeating statements in inner speech over the span of 30 seconds (e.g., “…That was stupid…That was stupid…”).

**

Figure 4. An illustration of the inner speech condition in Ford et al. (2001a)

The authors found that the auditory probes (but not the visual checkerboard) were attenuated in the inner speech condition relative to the baseline listening condition. Ford and Mathalon claim that their results confirm *Attenuation*: inner speech “may reflect the action of a corollary discharge [efference copy] from frontal brain structures preparing auditory cortex for speech” (p. 45).

Although both Whitford et al. and Ford et al. claim to support the association of inner speech with a prediction, their results are in tension with one another. Whitford et al. found attenuation of the incoming speech signal relative to listening only when there was a *match* between the speech sound produced in inner speech and the sound played over the headset. However, Ford et al. found attenuation relative to listening even when the sounds played over the headset – speech sounds and broadband sound – *did not match* the sentence produced in inner speech – “That was stupid”. What gives?

To understand these authors’ conflicting results, we must recognize that there are a number of ways the processing of an incoming signal can be attenuated. One cause of attenuation is repetition. Repeated presentations of a stimulus have been shown to lead to attenuated processing of that stimulus. Relative to the first presentation, there is less activation during the second presentation of the same stimulus. For example, when presented with two consecutive identical images of a campus scene, there is suppression of neural response in the parahippocampal place area for the second image compared to the first (Epstein et al., 2008). Crucially, this effect is also observed in mental imagery. For example, repeatedly imagining the same musical notes results in attenuated neural response (Navarro Cebrian and Janata, 2010).

Another cause of attenuation is expectation. Controlling for repetition, when a subject expects a stimulus, there is less neural activation during the presentation of that stimulus than when she does not expect it. Summerfield (2008) placed subjects in two conditions involving either the repetition of the same face or alternation of different faces. In the first condition, the repeated faces were more probable than the alternating faces, whereas in the second condition, the alternating faces were more probable than the repeated faces. The authors found that neural suppression in the fusiform face area was a function of the *probability* of repetition or alternation and not a function of brute repetition. Based on these results, a number of authors have concluded that there are at least two distinct mechanisms underlying attenuation: repetition of a stimulus and expectation of a stimulus (e.g., Todorovic and Lange, 2012). Using this insight, we can make sense of the apparent conflict between the Whitford et al. and Ford et al. results, but in a way that fails to support predictive models of inner speech.

Recall that Ford et al. had their subjects *repeat* a sentence in inner speech for 30 seconds while subjects were presented with probes that alternated between a visually presented checkerboard, speech sounds, and broadband noise. The difference between the control and experimental conditions is the repetition of the sentence in inner speech. We know that the attenuation of the incoming stimulus cannot be an effect of comparing the repeated sentence with the incoming stimulus, since there is a mismatch between the two. Given the difference between the conditions, the only other option is that the observed attenuation stems from the mere repetition of the sentence in inner speech. This hypothesis is confirmed by the evidence presented above that repeated imaginings are alone sufficient to generate attenuation. The attenuation is not caused by comparing inner speech and incoming stimulus, as would be required if inner speech were a prediction, but by the brute repetition of a sentence in inner speech.

Whitford et al.’s results fare no better. The relationship between prediction and data seems to be *graded*: the more of a match there is between prediction and data, the more the data are attenuated (Niziolek, et al., 2013). Thus, if inner speech is a prediction, we should expect more attenuation when one hears /bi/ and produces /ba/ in inner speech than (say) when one hears /fi/ and produces /ba/, since the former pair share more auditory features than the latter pair. However, Whitford et al. did not find a significant difference between the level of attenuation produced by the /ba/-/bi/ mismatch and the level of attenuation produced by simply listening to /bi/, despite the fact that hearing /ba/ while producing /bi/ in inner speech involves more overlap in auditory features than hearing /bi/ without inner speech. A more plausible explanation of the Whitford et al. finding is that it is the result of attenuation via expectation. Subjects were cued to produce either /ba/ or /bi/ at the outset of each run and the objective probability that the cued speech sound would be presented was .5. Within this unpredictable environment, I suggest that cueing subjects to produce a speech sound primes them to expect that speech sound to occur, resulting in the attenuation of the speech sound that matches the cue-based expectation.

There has also been an attempt to confirm *Attenuation* behaviorally. Scott et al. (2013) presented subjects with an ambiguous or hybrid speech sound between /ava/ and /aba/ while subjects concurrently produced /afa/ or /apa/ in inner speech.

 Figure 5. An illustration of the inner speech conditions in Scott et al. (2013)

The authors found that subjects tended to report perceiving the hybrid speech sound as /ava/ if they produced /afa/ and as /aba/ if they produced /apa/.Their explanation is that inner speech ‘captures’ the hybrid speech sound: the hybrid sound is heard as /ava/ or /aba/ depending on the character of the middle consonant produced in inner speech. Since /p/ shares more articulatory/auditory properties with /b/ than with /f/ (/p/ and /b/ are bilabial stops), producing /apa/ generates a perception of /aba/, and since /f/ shares more articulatory/auditory properties with /v/ than with /b/ (/f/ and /v/ are labiodental fricatives), producing /afa/ generates a perception of /ava/. Scott et al.’s basic idea is that inner speech serves as a prediction that pulls the hybrid sound in its direction, thereby generating percepts that are closer to it. They conclude that their result “provide[s] support for the claim that inner speech has rich auditory content...that is provided by corollary discharge [efference copy]” (p. 291).

However, this result actually directly contradicts *Attenuation*. Take the case where /afa/ is produced in inner speech and the hybrid is heard as /ava/. Why do subjects hear /ava/? Scott et al. seem to assume that there is some articulatory/auditory property that /afa/ shares with the hybrid that /aba/ does not share with the hybrid, such that when /afa/ is produced in inner speech, that common property is highlighted or enhanced and /ava/ is heard. The problem is that this common property should be *attenuated*, according to predictive models. When subjects concurrently produce /afa/ while being presented with the hybrid, there should be, from the perspective of predictive models, a greater probability of reporting *not* hearing /ava/. However, this prediction is directly contradicted by the result obtained by Scott et al. Therefore, the result does not support predictive models of inner speech. Instead, the result seems best explained by simple priming effects which cause an ambiguous stimulus to be interpreted in line with an imagined prime.

This sample of representative studies fail to validate *Attenuation*, and therefore fail to provide empirical support in favor of the Standard and Predictive Processing Models. At most, they reveal priming effects of various kinds: Ford et al. show repetition priming (repetition of an auditory stimulus leads to a decrease in auditory activity), Whitford et al. show expectation-based priming (an increase in the subjective probability that /ba/ will occur leads to a decrease in auditory activity when it does occur), and Scott et al. show perceptual priming (because /ava/ and /afa/ sound similar, the hybrid is heard as /ava/). Moreover, at least two of the studies suggest a stronger conclusion. The Ford et al. and Scott et al. results suggest that inner speech shows an effect – attenuation in the mismatch case – that indicates that inner speech is not a prediction.

I believe we can generalize this conclusion beyond the case of inner speech: that imagery in general is not a prediction. Whereas predictions make salient data that *conflict* with them, mental images make salient data that are *consistent* with them. Predictions make salient data that are *dissimilar* to them. Support for this comes from the mormyrid fish, a nocturnal fish that uses electrolocation to navigate and track prey. The fish sends an electric pulse into its environment, a copy of which is used to predict the incoming electric pulse were no creatures present in its environment. This prediction is subtracted from the total returning signal impinging on its electrosensory receptors. The portion of the returning signal that was predicted is not further processed, while the remaining portion is enhanced and made available for the tracking of prey (Bell, 1989). Salient information – information relevant to navigation and prey location – is thus represented in that portion of the incoming signal that fails to match the prediction. There is also more direct empirical evidence that mismatches between predictions and data determine salience. For example, predictable visual stimuli do not capture attention and so are easier to ignore than unpredictable stimuli (see, e.g., Itti and Baldi, 2009).

In contrast, mental images make salient data that are *similar* to them. For example, Farah (1985) found that when a subject is presented with a faint image of an “H” or “T”, the subject’s report of which letter she sees will tend to align with whether she is visually imagining an “H” or “T,” respectively. Relatedly, Pearson et al. (2008) find that the perception of a bistable image – an image with two fixed interpretations – is biased in the direction of the visual imagery one is producing. Similar findings have been shown using fMRI (Lu et al., 2017), and have also been discovered in the olfactory (Tomiczek and Stevenson, 2009) and auditory (Pitt and Crowder, 1992) modalities. The literature on imagery provides overwhelming evidence that imagery implicates such salience effects. Empirical research therefore suggests that imagery does not have predictive effects, and so is not a prediction.

One might object that the predictive processing framework does account for the idea that mental images make salient data that are similar to them. After all, predictive processing theorists have claimed to be able to explain many salience-type phenomena, including those found in binocular rivalry (Hohwy, Roepstorff, and Friston, 2008), the Posner paradigm (Feldman and Friston, 2010), and figure-ground discrimination (Kanai, et al., 2015). In attempting to extend this list, the predictive processing theorist may account for the salience effects of imagery by claiming that imagery is a prediction *with a high level of precision*. On this view, the high-level of precision accords the prediction a larger weight relative to lower-level predictions in the hierarchy, thereby forcing lower-level predictions to be updated in a manner that bring them in line with the prediction. Thus, for example, if a visual image of a bird is a high-precision prediction that *one sees a bird*, the visual image will bring lower-level predictions in line with it, thereby making salient bird-like features in one’s environment.

The problem with this explanation is that salient features are associated with attention to those features. According to the predictive processing framework, attention involves an *increase* of precision on prediction error (Feldman and Friston, 2010; Hohwy, 2013; Clark, 2015). However, as we noted in Section 3, the framework claims that imagery involves a *decrease* of precision on prediction error. There is thus a puzzle in the predictive processing model now under consideration. On the one hand, if the visual image is to have salience effects, one must *upmodulate* the precision of the prediction error generated by the prediction that *one sees a bird*. However, this is difficult to square with the claim that construction of the visual image involves *downmodulating* the precision of the prediction error generated by the prediction that *one sees a bird*. The burden thus remains with the predictive processing theorist to reconcile their account of imagery with the dominant predictive processing view of attention.

## 4. A Theoretical Problem with Predictive Processing Models of Imagery

I have so far criticized both predictive models on empirical grounds. In the present Section, I focus only on the (increasingly popular) predictive processing model. I will argue that the model faces a damning theoretical problem: if imagery is a means of minimizing prediction error, then perception and action become otiose. According to predictive processing models, what distinguishes imagery from both perception and action is that prediction error is not minimized via revising either the prediction or the data. Rather, prediction error is minimized via downmodulating the *precision* of the prediction error. Consider again how this is supposed to play out in inner speech. The Predictive Processing Model identifies inner speech with a high-precision prediction – *there are incoming speech sounds in my own voice*. Despite the fact that there is a difference between this prediction and the data – *there are no incoming speech sounds in my own voice* – this difference issues neither in speech perception nor speech production because, according to Wilkinson and Fernyhough, decreasing the precision on prediction error minimizes prediction error without revising the prediction (which would generate speech perception) or the data (which would generate speech production).

The problem with introducing this precision parameter is that downmodulation of the precision on prediction error offers the subject a standing resource to *always* avoid revising their prediction or data. With the ability to downmodulate precision, subjects have every reason to reject the life of perception and action in favor of the life of imagination: in conversation, I need not respond to my interlocutor, but only imagine responding; I need not see the obstacle in front of me, but only imagine seeing it; I need not drink, but only imagine the satisfaction of quenched thirst. Indeed, minimizing prediction error via revision of prediction or data faces far more potential obstacles and frustrations than minimizing prediction error via downmodulating the precision on prediction error. Revising one’s predictions to match incoming data and changing incoming data to match one’s predictions face routine upsets in the form of illusions and failed actions. However, it is not clear that simple downmodulation is prone to such potential upsets. Of course, downmodulating precision on prediction error to the exclusion of revising prediction or data would have disastrous consequences for the subject, leading to a swift death. I call this the ‘fantasy problem.’ The predictive processing theorist owes us a reason for thinking that the subject would not always downmodulate the precision term when faced with prediction error.

A response on behalf of the predictive processing theorist can be formulated once we note the similarity between the fantasy problem and the much-discussed ‘dark room problem’. The dark room problem underscores a family of issues related to how the predictive processing framework accounts for motivation (Sims, 2017). According to one version of the problem, the most effective way of minimizing prediction error is to remain in a perfectly predicted, unchanging environment – a so-called ‘dark room’ – until one dies. According to proponents of the dark room problem, since we in fact engage with a variety of highly unpredictable environments, the predictive processing framework cannot account for why we go about perceiving and acting.

Though there is a superficial similarity between the dark room problem and the fantasy problem, there is a deep and significant difference. The creature of the dark room and the creature of fantasy both enter a situation – the dark room or imagination – where they revise neither prediction nor data. However, the means by which the creatures achieve this are different: the creature of the dark room achieves zero prediction error by placing itself in a situation where prediction and data match, whereas the creature of fantasy achieves zero prediction error regardless of whether there is a match between prediction and data. In other words, while the creature of the dark room must find a place where it receives perfectly predicted inputs, the creature of fantasy need not do so, since its strategy of revision-avoidance is executable whatever its relation to the world happens to be.

This difference is key to showing how standard responses to the dark room problem do not transfer to the fantasy problem. In response to the dark room problem, predictive processing theorists often argue that evolution has shaped us such that we would in fact generate *high* prediction error if we remained in highly predictable states. For example, according to Hohwy (2013), what gets the creature out of the dark room is that it is equipped with a higher-level prediction that *there is a non-zero rate of prediction error*. Since there is no prediction error in the dark room, this higher-level prediction fails to match the incoming data – *there is no prediction error*. This mismatch, in turn, generates prediction error, thereby causing the creature to minimize it by generating lower-level predictions that fail to match incoming data, i.e., by ‘escaping’ the dark room (Hohwy, 2013).

However, this response fails to address the fantasy problem. Let us grant that the creature of fantasy is equipped with the higher-level prediction that *there is a non-zero rate of prediction-data revisions*. Of course, that prediction will generate prediction error for the creature of fantasy. But there is nothing to stop the creature of fantasy from downmodulating the precision on *this* higher-order prediction error just as it downmodulated the precision on first-order prediction errors. The higher-level prediction that *there is a non-zero rate of prediction-data revisions* will itself be folded into its imagination. It will not help at this point to claim that the higher-level prediction is a hyperprior, a higher-level prediction that is also impossible to revise. If the prediction that *there is a non-zero rate of prediction-data revisions* is a hyperprior, then it follows that the predictionis impossible to revise, but it does not follow that the precision of the resulting prediction error is impossible to downmodulate. [[3]](#footnote-3)

## 5. Sketching an Alternative: Toward a Goal-Based Model of Imagery

I have so far argued that predictive models of imagery fail to account for the salience effects of imagery (Section 3), and that predictive processing models of imagery undermine any motivation a subject may have to perceive the world and act in it (Section 4). Since predictive models constitute the most empirically sophisticated cognitive theory of mental imagery, we have an empirical and theoretical case against the view that imagery is a kind of cognitive state, a prediction about the world being a certain way. Of course, this argument alone does not entail that mental images are not cognitive states, since there are cognitive models of imagery that do not identify images with predictions.[[4]](#footnote-4) Nevertheless, the problems for predictive models stemming from salience and motivation highlight what I will call the ‘conative profile’ of imagery. In the current section, I will suggest that we should adopt a conative model of imagery in order to account for its conative profile.

Imagery has well-known effects on motivation and performance. For example, subjects who use mental imagery tend to engage in more golf (Martin and Hall, 1995), exercise more (Gammage, et al., 2000), and are more motivated to pursue a planned activity (Renner, et al., 2019). Imagery also leads to improvement in performance: subjects who engage in imagery have more effective tennis serves (Guillot, et al., 2013), see improved movement after stroke (Page et al., 2007), and are better able to execute novel actions (Mulder, et al., 2004). Of course, imagery’s effect on motivation and performance is a double-edged sword: smokers who engage in smoking-related imagery also have an increased urge to smoke (Tiffany and Hakenewerth, 1991), and imagery incongruent with one’s current task tends to hamper performance (Ramsey, et al., 2010). In general, an image of X tends to motivate in the direction of X by making one more likely to fulfill X (Nanay, 2020). How should we account for the conative profile of imagery? Not by appeal to prediction, or so I have argued in this paper.

What is the alternative? We can account for the motivational and performance effects of imagery, I suggest, if we treat imagery as a conative state, representing things as *to be made* a certain way. The conative states with which I will identity imagery are goal states, those sensory states that have the functional role of generating motor commands. I will first illustrate the account within the standard framework of motor control, returning to the example of inner speech (Figure 6).

 Figure 6. The Goal Model of inner speech

On the “Goal Model,” inner speech is the sensory representation of speech sounds. The sensory representation feeds into an inverse model, which transforms it into a set of speech motor commands. Thus, a sensory representation of a [p] sound will be fed into the inverse model, which will generate motor commands for {[labial, -round], [-voice], [+stop]}, which, if executed, will generate an actual [p] sound that matches the initial sensory representation of [p]. Unlike in speech production, however, in inner speech the inverse model is inhibited, thereby inhibiting the transformation of sensory into motor representation. Although inner speech is a sensory representation, what makes that representation a goal state is that it has the functional profile of a goal state, namely, that it issues in motor commands. An analogous account can be generated for motor (kinesthetic and proprioceptive) imagery as well. In general, on this view, imagery is a goal state in the sense that imagery is identical to the activation of a sensory representation with the functional role of a goal state, but where execution of the goal is suppressed. Mental images are conative, not cognitive: they are sensory representations that represent the world as to be made a certain way.

How does the goal model of imagery differ from standard and predictive processing models? The goal model draws on the same architecture as the standard model. The difference concerns where imagery is found in that architecture: while the standard model identifies it with a prediction, the goal model identifies it with a goal. The goal model disagrees with the predictive processing model about the representational primitives that compose the architecture. While the goal model appeals to such states as goal representations and motor commands, the predictive processing model tries to reduce these states to predictive processes. However, the fantasy problem suggests that this reductive project is not compatible with an explanation of imagery. We need to appeal to the architecture of the standard model, while treating imagery not as a prediction but as a goal.

What is the scope of this conative view of imagery? If we adopt the standard model of motor control, it may seem that the view only applies to motor-related forms of imagery, such as inner speech and motor imagery, and not to perceptual forms like visual imagery. This inference would be mistaken, however. Researchers have shown that visual imagery correlates with gaze pattern (Brandt and Stark, 1997; Laeng et al., 2014). For example, visually imagining the lower portion of an apple correlates with downward eye movements.[[5]](#footnote-5) These findings make visual imagery amenable to a goal-based analysis. According to this analysis, visual imagery is identical to a sensory representation of seeing an object from a perspective. In the case of normal eye movement, the sensory representation generates motor commands for eye movements, which, when executed, result in moving one’s eyes in a direction that brings about seeing an object from a perspective. In the case of visual imagery, however, the sensory representation of seeing an object from a perspective is activated but downstream processes are suppressed. The central idea is that the same conative processes that are used to drive motor commands relevant to perception and action are re-used in generating mental images. [[6]](#footnote-6)

One might object to the identification of images with goals by arguing that it is possible to conjure a visual image without having a goal the execution of which would satisfy the image. For example, I might have the goal of avoiding looking into my dark, scary basement, but nevertheless conjure up a visual image of the staircase leading into my basement. This objection loses traction once we recognize that imagery can be realized either by online *or* offline goal states. Although I may have an *online* goal to avoid looking down into the basement, nevertheless I conjure up a visual image of the basement by conjuring up an *offline* goal to look down into the basement. The latter state has the functional profile of a goal, but is taken offline and so does not lead to movement toward the basement. It might strike one as strange that generating imagery sometimes requires such contradictory goals, but, as we shall soon see, it is just this feature that puts us in a position to account for interference effects of imagery.

The goal-based model of imagery is able to explain at least three phenomena that defy explanation by predictive models, as discussed in Sections 3 and 4.

Recall that imagery can generate positive effects on performance. For example, mental imagery of a proper tennis serve leads to better tennis serves. In effect, fine-tuning one’s mental imagery leads to better performance. According to the goal model, properly fine-tuning one’s imagery just is a way of properly fine-tuning one’s goals. On the current view, then, the explanation of positive performance effects reduces to the explanation of why properly fine-tuned goals lead to positive performance. Of course, I have not provided an explanation of the latter fact, but that there is some such explanation should not be controversial.

Recall also that imagery makes salient features that are consistent with the image. According to the current view, salience via imagery is a species of salience via goals. Salience is thought to be a product of both a bottom-up, *stimulus-based* salience map and a top-down, *goal-based* salience map (Fecteau and Munoz, 2006). Salience maps are maps of the environment that provide measures of salience across features of the environment. The values of the bottom-up map are driven by stimuli in the environment, while the values of the top-down map are driven by goals one has with respect to the environment. On the current view, then, mental images help shape goal-based salience maps, which is what explains why features consistent with a mental image are salient.

Finally, imagery can also interfere with performance. For example, finding Waldo (target) in a group of people is hampered when one maintains a mental image of Whitebeard (distractor) (see Segal and Fusella (1970) for a similar result). The ingredients used to account for salience effects can also be used to account for such interference effects. Maintaining a mental image of Whitebeard increases the salience value of the distractor, thereby hampering visual detection of Waldo. Of course, whether these explanations of performance, salience, and interference effects are correct requires an assessment of empirical work that I do not have space for in this paper. However, one thing should be clear: the explanations provided by goal-based models of imagery present promising alternatives to traditional predictive models, which cannot account for the direction of salience effects of imagery.

One might object that if imagery is a goal taken *offline*, then imagery cannot have positive effects on performance nor shift salience values in a goal-based salience map. According to the objection, only an *online* goal state, one connected with the rest of the mental economy, can have such influences. My response is two-fold. First, this problem is not specific to the goal-based theory of imagery. The same type of problem arises for any theory of imagery that assumes that images are realized when a state type with a specific functional profile, e.g.,a perceptual state, is re-used and taken offline. Second, taking a state offline is not all-or-nothing. In taking a state offline, certain connections are suppressed while others are maintained, and those that are suppressed are never absolutely suppressed. The contention in this paper is that the effects of imagery on performance, salience, and bodily and behavioral states more generally, are made possible by the partial suppression of goal states.

## 6. Conclusion: From Cognition to Conation

There is a strong tendency to think of imagery as a cognitive mental state – as falling on the cognitive side of the cognitive-conative divide. Predictive models of imagery are one way of making this background assumption concrete. But predictive models of imagery have turned out to be untenable because they fail to account for the conative profile of imagery. I have argued that we can account for its conative profile if we view imagery as falling on the conative side of the cognitive-conative divide. On this picture, imagery is a conative mental state: it is one of the shapes that our goals take.[[7]](#footnote-7)

**References**

Behroozmand, Roozbeh, and Charles R. Larson. (2011). Error-Dependent Modulation of Speech-Induced Auditory Suppression for Pitch-Shifted Voice Feedback. *BMC Neuroscience* 12 (1): 54. <https://doi.org/10.1186/1471-2202-12-54>.

Bell, C. C. (1989). Sensory Coding and Corollary Discharge Effects in Mormyrid Electric Fish. *Journal of Experimental Biology* 146 (1): 229–53.

Brandt, S. A., and L. W. Stark. (1997). Spontaneous Eye Movements during Visual Imagery Reflect the Content of the Visual Scene. *Journal of Cognitive Neuroscience* 9 (1): 27–38. <https://doi.org/10.1162/jocn.1997.9.1.27>.

Carruthers, Peter. (2018). The Causes and Contents of Inner Speech. In *Inner Speech: New Voices*, edited by Peter Langland-Hassan and Agustin Vicente. Oxford University Press. <http://www.oxfordscholarship.com/view/10.1093/oso/9780198796640.001.0001/oso-9780198796640-chapter-2>.

Clark, Andy. (2013). The Many Faces of Precision (Replies to Commentaries on ‘Whatever next? Neural Prediction, Situated Agents, and the Future of Cognitive Science’). *Frontiers in Psychology* 4. <https://doi.org/10.3389/fpsyg.2013.00270>.

———. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Reprint edition. Oxford University Press.

———. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*(3), 521–534. <https://doi.org/10.1007/s11097-017-9525-z>

Currie, G., & Ravenscroft, I. (1997). Mental Simulation and Motor Imagery. *Philosophy of Science*, *64*(1), 161–180. <https://doi.org/10.1086/392541>

Doggett, T., & Egan, A. (2007). Wanting Things You Don’t Want: The Case for an Imaginative Analogue of Desire. *Philosopher’s Imprint*, *7*(9). <http://hdl.handle.net/2027/spo.3521354.0007.009>

Epstein, Russell A., Whitney E. Parker, and Alana M. Feiler. (2008). Two Kinds of FMRI Repetition Suppression? Evidence for Dissociable Neural Mechanisms. *Journal of Neurophysiology* 99 (6): 2877–86. <https://doi.org/10.1152/jn.90376.2008>.

Farah, Martha J. (1985). Psychophysical Evidence for a Shared Representational Medium for Mental Images and Percepts. *Journal of Experimental Psychology: General* 114 (1): 91–103. <https://doi.org/10.1037/0096-3445.114.1.91>.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences*, *10*(8), 382–390. <https://doi.org/10.1016/j.tics.2006.06.011>

Feldman, Harriet, and Karl J. Friston. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience* 4. <https://doi.org/10.3389/fnhum.2010.00215>.

Ford, J. M., D. H. Mathalon, S. Kalba, S. Whitfield, W. O. Faustman, and W. T. Roth. (2001a). Cortical Responsiveness during Talking and Listening in Schizophrenia: An Event-Related Brain Potential Study. *Biological Psychiatry* 50 (7): 540–49. <https://doi.org/10.1016/s0006-3223(01)01166-0>.

Ford, Judith M., Daniel H. Mathalon, Sontine Kalba, Susan Whitfield, William O. Faustman, and Walton T. Roth. (2001b). Cortical Responsiveness During Inner Speech in Schizophrenia: An Event-Related Potential Study. *American Journal of Psychiatry* 158 (11): 1914–16. <https://doi.org/10.1176/appi.ajp.158.11.1914>.

Ford, Judith M., and Daniel H. Mathalon. (2004). Electrophysiological Evidence of Corollary Discharge Dysfunction in Schizophrenia during Talking and Thinking. *Journal of Psychiatric Research* 38 (1): 37–46. <https://doi.org/10.1016/s0022-3956(03)00095-5>.

Freud, S., & Gay, P. (1989). *An Autobiographical Study* (J. Strachey, Ed.; The Standard edition). W. W. Norton & Company.

Friston, K., Thornton, C., & Clark, A. (2012). Free-Energy Minimization and the Dark-Room Problem. *Frontiers in Psychology*, *3*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00130>

Gammage, K. L., Hall, C. R., & Rodgers, W. M. (2000). More about Exercise Imagery. *The Sport Psychologist*, *14*(4), 348–359. <https://doi.org/10.1123/tsp.14.4.348>

Grandchamp, R., Rapin, L., Perrone-Bertolotti, M., Pichat, C., Haldin, C., Cousin, E., Lachaux, J.-P., Dohen, M., Perrier, P., Garnier, M., Baciu, M., & Lœvenbruck, H. (2019). The ConDialInt Model: Condensation, Dialogality, and Intentionality Dimensions of Inner Speech Within a Hierarchical Predictive Control Framework. *Frontiers in Psychology*, *10*, 2019. <https://doi.org/10.3389/fpsyg.2019.02019>

Guillot, A., Desliens, S., Rouyer, C., & Rogowski, I. (2013). Motor Imagery and Tennis Serve Performance: The External Focus Efficacy. *Journal of Sports Science & Medicine*, *12*(2), 332–338.

Heinks-Maldonado, Theda H., Daniel H. Mathalon, Max Gray, and Judith M. Ford. (2005). Fine-Tuning of Auditory Cortex during Speech Production. *Psychophysiology* 42 (2): 180–90. <https://doi.org/10.1111/j.1469-8986.2005.00272.x>.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. <https://doi.org/10.1016/j.cognition.2008.05.010>

Hohwy, Jakob. 2013. *The Predictive Mind*. 1 edition. OUP Oxford.

Itti, Laurent, and Pierre Baldi. (2009). Bayesian Surprise Attracts Human Attention. *Vision Research*, Visual Attention: Psychophysics, electrophysiology and neuroimaging, 49 (10): 1295–1306. <https://doi.org/10.1016/j.visres.2008.09.007>.

Jack, Bradley N., Mike E. Le Pelley, Nathan Han, Anthony W. F. Harris, Kevin M. Spencer, and Thomas J. Whitford. (2019). Inner Speech Is Accompanied by a Temporally-Precise and Content-Specific Corollary Discharge. *NeuroImage* 198: 170–80. <https://doi.org/10.1016/j.neuroimage.2019.04.038>.

Jeannerod, M. (2006). *Motor Cognition: What Actions Tell to the Self*. Oxford University Press.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140169. <https://doi.org/10.1098/rstb.2014.0169>

Laeng, Bruno, Ilona Bloem, Stefania D’Ascenzo, and Luca Tommasi. (2014). Scrutinizing Visual Images: The Role of Gaze in Mental Imagery and Memory. *Cognition* 131 (May): 263–283. <https://doi.org/10.1016/j.cognition.2014.01.003>.

Langland-Hassan, P. (2012). Pretense, imagination, and belief: The Single Attitude theory. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *159*(2), 155–179.

Lœvenbruck, H., R. Grandchamp, L. Rapin, L. Nalborczyk, M. Dohen, P. Perrier, M. Baciu, and M. Perrone-Bertolotti. (2018). A Cognitive Neuroscience View of Inner Language: To Predict and to Hear, See, Feel. In *Inner Speech: New Voices*, edited by Peter Langland-Hassan and Agustin Vicente. Oxford University Press. <http://www.oxfordscholarship.com/view/10.1093/oso/9780198796640.001.0001/oso-9780198796640-chapter-6>.

Lu, Lingxi, Changxin Zhang, and Liang Li. (2017). Mental Imagery of Face Enhances Face-Sensitive Event-Related Potentials to Ambiguous Visual Stimuli. *Biological Psychology* 129 (October): 16–24. <https://doi.org/10.1016/j.biopsycho.2017.07.013>.

Mainland, J., & Sobel, N. (2006). The sniff is part of the olfactory percept. *Chemical Senses*, *31*(2), 181–196. <https://doi.org/10.1093/chemse/bjj012>

Martin, K. A., & Hall, C. R. (1995). Using mental imagery to enhance intrinsic motivation. *Journal of Sport & Exercise Psychology*, *17*(1), 54–69.

Mulder, T., Zijlstra, S., Zijlstra, W., & Hochstenbach, J. (2004). The role of motor imagery in learning a totally novel movement. *Experimental Brain Research*, *154*(2), 211–217. <https://doi.org/10.1007/s00221-003-1647-6>

Myers, J. (2022). Imaginative Beliefs. [Unpublished Manuscript].

Nanay, B. (2020). Motor Imagery and Action Execution. *Ergo, an Open Access Journal of Philosophy*, *7*. <https://doi.org/10.3998/ergo.12405314.0007.013>

Navarro Cebrian, A., & Janata, P. (2010). Electrophysiological correlates of accurate mental image formation in auditory perception and imagery tasks. *Brain Research*, *1342*, 39–54. <https://doi.org/10.1016/j.brainres.2010.04.026>

Niziolek, C. A., S. S. Nagarajan, and J. F. Houde. (2013). What Does Motor Efference Copy Represent? Evidence from Speech Production. *Journal of Neuroscience* 33 (41): 16110–16.

Page, S. J., Levine, P., & Leonard, A. (2007). Mental practice in chronic stroke: Results of a randomized, placebo-controlled trial. *Stroke*, *38*(4), 1293–1297. <https://doi.org/10.1161/01.STR.0000260205.67348.2b>

Patel, S. (2021). From speech to voice: On the content of inner speech. *Synthese*, *199*(3), 10929-10952. <https://doi.org/10.1007/s11229-021-03274-6>

Pearson, Joel, Colin W. G. Clifford, and Frank Tong. (2008). The Functional Impact of Mental Imagery on Conscious Perception. *Current Biology: CB* 18 (13): 982–86. <https://doi.org/10.1016/j.cub.2008.05.048>.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences*, *36*(4), 329–347. <https://doi.org/10.1017/S0140525X12001495>

Pitt, M. A., and R. G. Crowder. (1992). The Role of Spectral and Dynamic Cues in Imagery for Musical Timbre. *Journal of Experimental Psychology. Human Perception and Performance* 18 (3): 728–38. [https://doi.org/10.1037//0096-1523.18.3.728](https://doi.org/10.1037/0096-1523.18.3.728).

Ramsey, R., Cumming, J., Eastough, D., & Edwards, M. G. (2010). Incongruent imagery interferes with action initiation. *Brain and Cognition*, *74*(3), 249–254. <https://doi.org/10.1016/j.bandc.2010.08.005>

Renner, F., Murphy, F. C., Ji, J. L., Manly, T., & Holmes, E. A. (2019). Mental imagery as a “motivational amplifier” to promote activities. *Behaviour Research and Therapy*, *114*, 51–59. <https://doi.org/10.1016/j.brat.2019.02.002>

Scott, Mark, H. Henny Yeung, Bryan Gick, and Janet F. Werker. (2013). Inner Speech Captures the Perception of External Speech. *The Journal of the Acoustical Society of America* 133 (4): EL286-292. <https://doi.org/10.1121/1.4794932>

Segal, S. J., & Fusella, V. (1970). Influence of imaged pictures and sounds on detection of visual and auditory signals. *Journal of Experimental Psychology*, *83*(3, Pt.1), 458–464. <https://doi.org/10.1037/h0028840>

Shah, N., & Velleman, J. D. (2005). Doxastic Deliberation. *The Philosophical Review*, *114*(4), 497–534.

Sims, Andrew. (2017). The Problems with Prediction. *PPP – Philosophy and Predictive Processing*, edited by Thomas Metzinger and Wanja Wiese. <https://predictive-mind.net/>

Sprevak, Mark (2021). *Predictive coding IV: The implementation level.* [Preprint]

Summerfield, Christopher, Jim M.P. Monti, Emily H. Trittschuh, M.-Marsel Mesulam, and Tobias Egner. (2008). Neural Repetition Suppression Reflects Fulfilled Perceptual Expectations. *Nature Neuroscience* 11 (9): 1004–6.

Swiney, Lauren, and Paulo Sousa. (2014). A New Comparator Account of Auditory Verbal Hallucinations: How Motor Prediction Can Plausibly Contribute to the Sense of Agency for Inner Speech. *Frontiers in Human Neuroscience* 8: 675. <https://doi.org/10.3389/fnhum.2014.00675>.

Tian, Xing, and David Poeppel. (2010). Mental Imagery of Speech and Movement Implicates the Dynamics of Internal Forward Models. *Frontiers in Psychology* 1: 166. <https://doi.org/10.3389/fpsyg.2010.00166>.

Tiffany, S. T., & Hakenewerth, D. M. (1991). The production of smoking urges through an imagery manipulation: Psychophysiological and verbal manifestations. *Addictive Behaviors*, *16*(6), 389–400. <https://doi.org/10.1016/0306-4603(91)90047-L>

Todorovic, A., & de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(39), 13389–13395. <https://doi.org/10.1523/JNEUROSCI.2227-12.2012>

Tomiczek, Caroline, and Richard Stevenson. (2009). Olfactory Imagery and Repetition Priming: The Effect of Odor Naming and Imagery Ability. *Experimental Psychology* 56 (February): 397–408. <https://doi.org/10.1027/1618-3169.56.6.397>.

Van Leeuwen, N. (2011). Imagination is where the action is. *The Journal of Philosophy*, *108*(2), 55–77.

Vicente, A., & Martínez-Manrique, F. (2016). The nature of unsymbolized thinking. *Philosophical Explorations*, *19*(2), 173–187. <https://doi.org/10.1080/13869795.2016.1176234>

Vygotsky, L. S. (1990). Imagination and Creativity in Childhood. *Soviet Psychology*, *28*(1), 84–96. <https://doi.org/10.2753/RPO1061-0405280184>

Whitford, Thomas J, Bradley N Jack, Daniel Pearson, Oren Griffiths, David Luque, Anthony WF Harris, Kevin M Spencer, and Mike E Le Pelley. (2017). Neurophysiological Evidence of Efference Copies to Inner Speech. Edited by Richard Ivry. *ELife* 6 (December): e28197. <https://doi.org/10.7554/eLife.28197>.

Wilkinson, Sam. (2014). Accounting for the Phenomenology and Varieties of Auditory Verbal Hallucination within a Predictive Processing Framework. *Consciousness and Cognition* 30 (November): 142–55. <https://doi.org/10.1016/j.concog.2014.09.002>.

Wilkinson, Sam, and Charles Fernyhough. (2017). Auditory Verbal Hallucinations and Inner Speech : A Predictive Processing Perspective. In *Before Consciousness : In Search of the Fundamentals of Mind.*, edited by Zdravko Radman, 285–304. Exeter: Imprint Academic. <http://books.imprint.co.uk/book/?gcoi=71157100617220>.

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, *11*(18), R729–R732. <https://doi.org/10.1016/S0960-9822(01)00432-8>

Ylinen, Sari, Anni Nora, Alina Leminen, Tero Hakala, Minna Huotilainen, Yury Shtyrov, Jyrki P. Mäkelä, and Elisabet Service. (2015). Two Distinct Auditory-Motor Circuits for Monitoring Speech Production as Revealed by Content-Specific Suppression of Auditory Cortex. *Cerebral Cortex* 25 (6): 1576–86. <https://doi.org/10.1093/cercor/bht351>.

1. Predictive processing theorists attempt to reduce conative process to predictive processes. One of the lessons of this paper is that a plausible account of imagery requires that conative processes be irreducible to predictive processes. [↑](#footnote-ref-1)
2. There remains much to be said concerning the specific cognitive and neural mechanisms that implement predictive processes (see, e.g., Sprevak (2021)). However, the arguments of this paper turn only on the abstract specification of predictive processing given here and not on the details of its implementation. [↑](#footnote-ref-2)
3. Clark (2018) offers a similar response to the dark room problem as Hohwy, claiming there is “something right” about responding to the dark room problem by appealing to types of prediction that mandate “change, exploration, and search” (p. 526). Clark is in general agreement with Friston’s claim that “[a]gents that predict rich stimulating environments will find the ‘dark room’ surprising and will leave at the earliest opportunity” (Friston, Thornton, and Clark, 2012, p. 3). The difference between Clark and Hohwy is just that Clark appeals to first-order predictions, while Hohwy appeals to higher-order predictions. Importantly, Clark (2018) claims that this kind of response remains “underspecified” without details about how such predictions are underwritten by interoceptive predictions and the assignment of precision by sub-cortical structures (p. 526-527). However, even if Clark’s response were fully fleshed out in terms of interoceptive predictions and sub-cortical structure, it is not clear that an analogous response to the fantasyproblem would succeed for the same reasons discussed here. The creature of fantasy can still downmodulate the precision on prediction error of the interoceptive predictions. [↑](#footnote-ref-3)
4. An alternative view holds that at least some images are beliefs (see, e.g., Langland-Hassan (2012) and Myers (2022)). In contrast to predictive models, extant belief-based models have yet to be spelled out in a way that make them amenable to empirical assessment. A general assessment of cognitive models of imagery thus awaits an empirical specification of belief-based models. [↑](#footnote-ref-4)
5. Data similar to that concerning visual imagery and eye movement seem to exist for the relationship between olfactory imagery and sniffing (Mainland and Sobel, 2006). [↑](#footnote-ref-5)
6. For other views that also link motivation and imagery see Freud (1989) and Vygotsky (1990). [↑](#footnote-ref-6)
7. Many thanks to Edouard Machery, Peter Langland-Hassan, Wayne Wu, James Shaw, Mark Wilson, Zina Ward, and Josh Myers for discussion and feedback on this paper. [↑](#footnote-ref-7)