Active inference, selective attention, and the cocktail party problem

Emma Holmes, Thomas Parr, Timothy D. Griffiths, Karl J. Friston

PII: S0149-7634(21)00420-6

DOI: https://doi.org/10.1016/j.neubiorev.2021.09.038

Reference: NBR 4362

To appear in: *Neuroscience and Biobehavioral Reviews*

Received Date: 12 April 2021

Revised Date: 27 August 2021

Accepted Date: 17 September 2021

Please cite this article as: Holmes E, Parr T, Griffiths TD, Friston KJ, Active inference, selective attention, and the cocktail party problem, *Neuroscience and Biobehavioral Reviews* (2021), doi: https://doi.org/10.1016/j.neubiorev.2021.09.038

Active inference, selective attention, and the cocktail party problem

Emma Holmes[1,2], Thomas Parr [2], Timothy D. Griffiths[2,3], and Karl J. Friston[2]

[1] *Department of Speech Hearing and Phonetic Sciences, UCL, London, WC1N 1PF, U.K.*

[2] *Wellcome Centre for Human Neuroimaging, UCL, London, WC1N 3AR, U.K.*

[3] *Biosciences Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, U.K.*

Corresponding author: Emma Holmes; E-mail: emma.holmes@ucl.ac.uk; Phone: +44 7597 967397; Mailing address: Department of Speech Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, U.K.

Abbreviated title: Active inference and selective attention

Highlights

- New generative model for selective attention during cocktail party listening
- Computational 'lesions' in the model dissociate different errors during word report
- We model different temporal hypotheses for preparatory attention
- Temporal changes in precision are necessary to explain ERPs but not reaction times
- CNV-like responses can be explained by subjective precision rather than action

Abstract

In this paper, we introduce a new generative model for an active inference account of preparatory and selective attention, in the context of a classic 'cocktail party' paradigm. In this setup, pairs of words are presented simultaneously to the left and right ears and an instructive spatial cue directs attention to the left or right. We use this generative model to test competing hypotheses about the way that human listeners direct preparatory and selective attention. We show that assigning low precision to words at attended—relative to unattended—locations can explain why a listener reports words from a competing sentence. Under this model, temporal changes in sensory precision were not needed to account for faster reaction times with longer cue-target intervals, but were necessary to explain ramping effects on event-related potentials (ERPs)—resembling the contingent negative variation (CNV)—during the preparatory interval. These simulations reveal that different processes are likely to underlie the improvement in reaction times and the ramping of ERPs that are associated with spatial cueing.

Keywords: Selective attention; Preparatory attention; Spatial attention; Temporal attention; Cocktail party listening; Active inference

Introduction

For decades, researchers and theoreticians have debated the basis of selective attention. A feature of attention is that different stimuli can be selected, without any changes in sensory signals (e.g., Driver, 2001; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999; Necker, 1832; Rubin, 1915; Van Noorden, 1975). In many situations, attention is not all-or-none, but rather fluctuates over time, even when the attribute that is attended is not obviously time itself (e.g., following a cue for location; Posner & Cohen, 1984; Rolke & Hofmann, 2007). In this paper, we examine the computational processes that can explain the implicit time-sensitive attentional set.

Time-dependent changes in attention can be understood at different temporal scales. Here, we focus on the gradual build-up of preparatory (i.e., anticipatory) attention, which occurs approximately 0–2 seconds before a target stimulus occurs (e.g., Holmes et al., 2018). Within this temporal window, a graded improvement in accuracy or response times occurs when an attentional cue is presented earlier in advance of an upcoming stimulus (Holmes, Kitterick, & Summerfield, 2018; Lu et al., 2009; Yamaguchi, Tsuchiya, & Kobayashi, 1994). There is also an anticipatory ramping of event-related potentials resembling the contingent negative variation (CNV) (Poljac & Yeung, 2012; Tecce, 1972; Walter, Cooper, Aldridge, McCallum, & Winter, 1964). These slow attentional effects can be distinguished from faster effects that occur at a theta rate (see Fiebelkorn & Kastner, 2019; Helfrich et al., 2018; Kotz, Schwartze, & Schmidt-kassow, 2009; Lakatos et al., 2016), which may be relevant to the temporal scheduling of stimulus sampling (e.g., rates of syllable presentation or the timing of saccades; Friston, Adams, Perrinet, & Breakspear, 2012; Giraud & Poeppel, 2012). Here, we focus on the mechanisms underlying slow changes, which have received less investigation. This paper offers a novel perspective, by introducing an active inference account of covert endogenous attention that we compare to human behaviour.

The requisite generative model is based on an influential paradigm introduced by Cherry (1953), termed 'cocktail party listening'. The original paradigm presents two semantically meaningful passages of speech to the left and right ears, and the listener is asked to shadow words in one ear. Subsequent follow-up studies have used simplified versions of this paradigm to isolate specific processes. For example, rather than passages of text, individual words, phrases, or sentences are presented—which may be devoid of semantic context and which listeners can report with a single response (e.g., a button press). These tasks are interesting for several reasons. First, different types of errors reflect distinct processes (Holmes et al., 2018; Johnsrude et al., 2013; Maddox & Shinn-Cunningham, 2012): errors can be words from the competing speech, a mixture of words from the target and competing speech, or words that were not spoken. When a listener reports words from the competing speech, this implies a problem with attentional selection: streams of words are correctly separated, but the incorrect talker is selected; whereas, reporting a mixture of words from the target and competing sentences implies that the two streams of words are not correctly separated.

Second, attentional preparation for a target talker appears to be graded: when a spatial cue (left or right) indicates which talker should be attended, reaction times progressively improve when the cue is presented longer before the talkers begin to speak (Holmes et al., 2018). Also, before the target talker begins to speak, electrophysiological responses resembling the CNV slowly increase over time (Holmes, Kitterick, & Summerfield, 2016, 2017), possibly reflecting activity in fronto-parietal brain regions (Hill & Miller, 2010; Lee et al., 2013).

In this paper, we examine how different computational processes affect accuracy, error types, reaction times, and electrophysiological responses in a simplified cocktail party paradigm, in which two pairs of words are presented at different locations. We simulate active inference under different generative models to test competing hypotheses about how these behavioural and electrophysiological hallmarks of preparatory attention emerge.

Crucially, the current paper focusses on the selective attention component of cocktail party listening. Other models have considered the extraction of words from an acoustic signal (Bornkessel-Schlesewsky, Schlesewsky, Small, & Rauschecker, 2015; Friston et al., 2021), acoustic segregation of a sound mixture (Chen, Luo, & Mesgarani, 2017), or whether stimuli are perceived as one or two streams (see Szabó, Denham, & Winkler, 2016 for a review). In contrast, the current paper focuses on the deployment of top-down attention. Unlike some other 'models' of attention, which are largely conceptual, here our goal is to make quantitative predictions and evaluate these based on empirical data.

We use active inference (Friston, 2003; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017) to simulate synthetic agents that act under different generative models. Active inference has been applied to a variety of topics in neuroscience (e.g., Brown, Adams, Parees, Edwards, & Friston, 2013; Brown, Friston, & Bestmann, 2011; Friston, 2005; Friston, Kilner, & Harrison, 2006; Mirza, Adams, Mathys, & Friston, 2016; Parr & Friston, 2017). It is based on the idea that perception uses approximate Bayesian inference, and the resulting behaviour can be cast as planning as inference. Under our generative model, beliefs about unobservable states of the world (e.g., the beliefs about the target words) determine observable outcomes (e.g., the visual cue that is presented and the words that are spoken at the two locations); inverting the model allows inferences about hidden states given sensory observations. Bayesian inference can be cast as minimising variational free energy: to update beliefs about hidden states to provide (simultaneously) the simplest and most accurate explanation of sensory data (see Equation A.2).

The enactive component of active inference enables a synthetic agent to act upon the world to change sensory observations. In this setting, beliefs about the best actions to take depend on expected outcomes, and can be formalised as minimising risk and ambiguity (see Equation A.3). Actions are more likely to be selected if they (i) generate preferred outcomes (i.e., they minimise risk: technically, the expected divergence between predicted and preferred outcomes) and (ii) if they reduce ambiguity (technically, the inverse precision of the likelihoods of particular outcomes, given beliefs about hidden states). Under this framework, we can simulate performance (i.e., action selection) and belief updating during perceptual inference under different generative models. By appealing to the accompanying process theories, electroencephalographic (EEG) responses can be modelled as consequent on the neuronal dynamics that minimise the energy (Friston, FitzGerald, et al., 2017). Unlike many other models, which either ignore or focus solely on electrophysiological responses, active inference allows us to simulate EEG and behavioural responses together under a variety of generative models.

Unlike other conceptualisations of attention, under models employing active inference, attention is conceptualised not as *what* is being represented, but rather as the *precision* of probabilistic representations (Feldman and Friston, 2010). Formally, this is specified as optimising the precision of mappings between hidden causes and sensory observations in the generative model (Hohwy, 2012; Parr & Friston, 2019b). This idea has been used to successfully model goal-dependent behaviour—for example, due to a change in task—that differs under the same sensory observations (Feldman & Friston, 2010; Mirza, Adams, Friston, & Parr, 2019; Parr, Corcoran, Friston, & Hohwy, 2019; Parr & Friston, 2017).

This paper aims to address two questions. First, can precision in different parts of a generative model explain the distinct ways in which attention can 'fail' in cocktail party listening—and thereby generate different types of errors? To study this, we apply previous formulations of attention in active inference to a simplified cocktail party listening paradigm, and consider how variations in precision could explain attentional effects that are not ‘all-or-none’. We use the ensuing model to test hypotheses about how precision in different parts of a generative model produces different types of error. Second, we turn to preparatory attention, which requires us to consider the temporal deployment of attention and attention. We ask whether behavioural and electrophysiological correlates of preparatory attention arise from similar processes. To study this, we extend previous work by considering precision as a state-dependent parameter that changes over time during preparatory attention—which has not been addressed previously. In our cocktail party listening paradigm, this reflects the time between the presentation of a spatial cue and the onset of a target phrase. Our goal was to make predictions about reaction times (RTs) and electroencephalographic (EEG) responses, and compare these predictions to empirical data reported by Holmes et al. (2018) and Holmes et al. (2017), respectively. Cucially, both of these studies employed a similar paradigm to study preparatory attention.

A generative model of selective attention during cocktail party listening

We first describe a generative model of cocktail party listening, and show that an agent who employs this model during active inference can solve the cocktail party problem. This provides the basis for testing hypotheses about how precision in different parts of the generative model affects performance, which we address in the next section of the paper.

Our generative model uses a (partially observable) Markov decision process (MDP) to map from discrete states to discrete outcomes. MDP models have been used in applications of active inference to many domains of neuroscience (e.g., Mirza et al., 2019, 2016; Parr & Friston, 2019a). In brief, MDP models describe how states change over time, and how states generate outcomes. Outcomes are observable features of the world, and states are hypotheses about the (hidden or latent) causes of outcomes. While states are not cognitive constructs themselves (see Ramstead, Friston & Hipólito, 2020), it can be helpful to name them in a manner that is interpretable. When we describe the states in this particular generative model, it should become clear that states are constructs that we assume people make inferences about in order to complete the task at hand. For example, the colour of an object is something that people infer from the signals that reach their visual system: it is not a directly observable property of the world, and is instead affected by the types of cone cells in the person’s retina and the lighting conditions in the room. Previous work (e.g., Feldman and Friston, 2010) has proposed that attention can be linked to the precision of the probability distributions that map between states and outcomes. These types of models are termed decision processes, because state transitions depend upon the decisions made (or policies selected). By inverting the MDP model, a synthetic subject can infer the most likely causes of sensory observations under their generative model, and infer the most probable course of action given those beliefs.

< FIGURES 1 & 2 >

Figure 1 depicts a simplified version of the cocktail party paradigm—resembling that used in many empirical studies—which consists of short phrases with closed-set responses: two talkers each speak a colour word and a number word (as in the Coordinate Response Measure corpus: Moore, 1981); the listener is directed to attend to the words spoken by the talker on their left or right by a visual cue (left or right arrow). There are four options for each colour word (“red”, “green”, “blue”, and “white”) and four options for each number word (“1”, “2”, “3”, “4”): on each trial, the listener is required to respond with the target colour and number words (which can be considered as analogous to pressing a button labelled “red 1”, as in empirical studies by Holmes et al., 2016, 2017, 2018 and others). While the paradigm used by Cherry (1953) used longer passages of speech that were semantically meaningful, the use of shorter phrases, devoid of semantic context allows us to isolate spatial attention. In this section, we begin with a simplified version of the task, which we extend later in the paper.

Figure 2 illustrates a plausible generative model for this task. The generative model contains four hidden state factors, and six outcome factors. The *Spatial Attention* factor contains two states, reflecting beliefs that attention is directed to the Left or Right. The *Target Colour* factor contains four states: the target word is Red, Green, Blue, or White. The *Target Number* factor also contains four states: the target word is 1, 2, 3, or 4. The *Response* factor contains 17 control states, which indicate the model's beliefs about the response emitted: 16 of these are the permissible combinations of the 4 colour and 4 number words, and the final state is Null, which means that no response is currently being made (e.g., at the beginning of a trial, before the correct response is inferred). The *Visual Cue* outcome factor indicates whether a Left or Right arrow is displayed on the screen. The *Left Colour Word* and *Right Colour Word* outcome factors indicate whether the left and right talkers, respectively, are saying the words Red, Green, Blue, or White. The *Left Number Word* and *Right Number Word* outcome factors indicate the same for the number word (1, 2, 3, or 4). The *Feedback* outcome factor describes the visual feedback that is observed: indicating whether the response is Correct or Incorrect. In summary, the generative model covers all the states that are necessary to generate observable outcomes that constitute the task or paradigm at hand.

The generative model takes the form specified in Appendix 2. Specifying the generative model formally requires the following probability distributions: (1) The likelihood mappings between hidden states and outcomes; (2) How states within each state factor are expected to transition over time; (3) Preferences over outcomes for each outcome factor; (4) Prior beliefs about which states the model is likely to be in, initially. Each probability distribution is specified in matrix form (see Equation A.4).

The likelihood matrix (A-matrix) for each outcome factor defines the mappings between each state and the possible outcomes. In other words, given that I believe that a particular word has been spoken, how likely am I to hear that word? Or, given I believe I am attending to the left, how likely am I to observe a left spatial cue? These examples may seem counterintuitive, because we are used to taking the perspective of perception: inferring states from observations in the world. From the perspective of perceptual inference (i.e., inferring the cause from its consequence), these mappings can be conceived as the model's confidence that the target word is Red, given that it has heard the word Red, and its confidence that it is attending to the left side, given it has seen a left spatial cue. We have set the parameters so that the model is very confident that the *Visual Cue* outcome will match its beliefs about which *Spatial Attention* state it is in. This is specified as an identity matrix and reflects an infinitely precise mapping. The *Visual Cue* outcome is not influenced by any of the other state factors.

The *Left Colour Word* outcome matches the model's beliefs about the *Target Colour* with infinite precision if the *Spatial Attention* state is Left, but there is no relationship between the *Target Colour* states and the *Left Colour Word* outcomes if the *Spatial Attention* state is Right: in the latter case, the likelihood mappings are specified as a uniform matrix (i.e., a matrix with zero precision). Similarly, the *Right Colour Word* outcome matches the model's beliefs about the *Target Colour* with infinite precision if the *Spatial Attention* state is Right, but with zero precision if it is Left. Equivalent mappings apply to the *Target Number*, *Left Number Word*, and *Right Number Word*. The *Feedback* outcome is Correct if the *Response* state matches the combination of the *Target Colour* and *Target Number* states (e.g., the *Target Colour* is Red, the *Target Number* is 1, and the *Response* is Red-1) and is Incorrect if it does not match. For this reason, the Null *Response* state is always considered Incorrect, regardless of the *Target Colour* and *Target Number* states.

A probably transition matrix (B-matrix) is specified for each state factor, and defines how states are expected to change over time. In other words, given I currently believe I'm attending to the left, how likely am I to believe I'm attending to the left at the next time step? In the current generative model, beliefs about *Spatial Attention*, *Target Colour* and *Target Number* states are stable, with infinite precision (that is, they are not expected to change over time, within a trial). The states within the *Response* factor are control states: in other words, an action (analogous to pressing a button) controls the *Response* state. The model contains 16 policies, which correspond to acting on the 16 permissible colour-number combinations. If the *Response* state is Null, then the action changes the *Response* state to the selected word with 100% probability. Otherwise, the action has no effect and the *Response* state remains the same, with infinite precision. In other words, once the synthetic agent has executed an action, it cannot change its mind. Specifying 16 policies, that correspond to the permissible colour-number combinations, means that a simulated agent acting under this generative model must choose a colour-number response on every trial, even if it is uncertain. Note that the Null response state is only used at the beginning of the trial, before the correct response is inferred.

The prior preference matrix (C-matrix) defines preferences over outcomes for each outcome factor. The current generative model has no preferences for seeing particular spatial cues (*Visual Cue* outcomes) or hearing particular words (*Left Colour Word*, *Right Colour Word*, *Left Number Word*, and *Right Number Word* outcomes). However, the model has a strong preference for Correct over Incorrect feedback (*Feedback* factor) (see Appendix 3 for parameter values).

The priors (D-matrix) define prior beliefs about which states the model is likely to be in, at the beginning of a trial. In the current model, we specified uniform beliefs over *Spatial Attention*, *Target Colour* and *Target Number* states—in other words, no cues or words were expected to be more likely than others *a priori*. However, the model has a strong (infinitely precise) prior belief that it is initially making no response (Null *Response* state). This allows the agent's action to affect the *Response* state once per trial.

Now that we have specified a plausible generative model, we can evaluate the performance of a synthetic agent that uses this generative model for active inference. In these simulations, the visual cue and the colour and number words for the two talkers are presented concurrently for simplicity; thus, each trial contains one round of belief updates (i.e., a one-shot trial), which includes updating beliefs about the *Response*, which is enacted. Given the model's prior belief that the *Response* is initially Null, the *Feedback* outcome always begins as Incorrect—in other words, the synthetic agent initially sees a cross on the screen. The preference for *Correct* feedback means that the agent's goal is to end the trial with a tick on the screen, by emitting a *Response* that it believes is correct. It can only do this by using the visual cue to infer which side it should be attending to, and using the colour and number words at that location to determine the correct response.

To check the task was performed accurately under this generative model, we simulated 100 trials using standard routines in SPM12 (using MATLAB R2017b). For each trial, we compared the response made by the agent with the sensory observations: if the action was consistent with the words at the target location (i.e., the *Left Colour Word* and *Left Number Word* if a Left *Visual Cue* was observed, and the *Right Colour Word* and *Right Number Word* if a Right *Visual Cue* was observed), the agent was considered to have performed the trial correctly and was awarded a point. Otherwise, it was considered to have performed incorrectly and received no points for that trial.

Happily, the synthetic agent performed 100% correct on this task (i.e., it made no errors). This serves as a proof of principle that the generative model described above enables perfect performance on this kind of task. In the next section, we 'break' (i.e., change the parameters of) different parts of the model, and assess how this affects task performance. Our goal is to simulate the types of errors made by human listeners.

Modelling impaired performance

Cocktail party listening is difficult for human listeners: Not only does competing speech mask the spectro-temporal components of target speech (“energetic masking”), but it also requires listeners to segregate target speech from a similar-sounding distracter (“informational masking”) (Brungart, 2001; Brungart, Simpson, Ericson, & Scott, 2001). Human errors are more likely to consist of words spoken by a competing talker than other words not spoken by any of the talkers, showing that informational masking plays a large role when cocktail party listening is challenging (Best, Ozmeral, Kopčo, & Shinn-Cunningham, 2008; Brungart & Simpson, 2002; Holmes et al., 2018). In this section, we ask which computational processes could lead to informational masking using the generative model introduced in the previous section. In other words, we changed the parameters of the generative model and examined their effects on percent correct and the types of errors that the agent made.

Three types of error have received interest in the literature (Holmes et al., 2018; Ihlefeld & Shinn-Cunningham, 2008; Johnsrude et al., 2013; Maddox & Shinn-Cunningham, 2012): “Masker errors”, in which reported words were spoken by the competing talker; “Mix errors”, in which some of the words were spoken by the target talker and others were spoken by the competing talker; and “Absent errors”, in which the words reported were spoken by neither talker. Distinguishing these errors is of interest, because they speak to different cognitive processes. In order to report the correct words in the paradigm used here, the two words from the target need to be correctly grouped together, which relies on streaming and, in addition, the correct talker needs to be attributed as the target. Masker errors suggest that two talkers have been streamed (i.e., the two consecutive words spoken by each talker have been successfully attributed to the same talker), but the listener failed to attend to the correct (i.e., target) talker. For example, this could arise if the target location cannot be distinguished from the competing location, or if the competing talker is salient and cannot be ignored. Mix errors suggest that the two talkers have not been successfully streamed, which could arise if the listener cannot infer which words belong to which talker. Absent errors could arise from failures in a variety of processes—including errors at the response stage, or low arousal or motivation. We, therefore, predicted that these different types of errors could be dissociated by changing different parameters of the generative model.

Reflecting the wide variety of processes that could lead to Absent errors, they are straightforward to simulate under our generative model. Absent errors arise if preferences over outcomes (C-matrix) are specified so that the agent no longer prefers correct feedback, because the agent has no reason to respond correctly. Alternatively, Absent errors could arise from stochastic selection of an action from posterior beliefs. Operationally, this stochasticity is parameterised with a parameter alpha: if alpha is less than one, the precision over the set of policies is low and the probability of selecting the least likely policy approaches the probability of selecting the most likely policy. Absent errors could also arise from imprecise beliefs about the relationship between the *Response* states and *Feedback* outcomes: In other words, this instils a belief that feedback is likely to be incorrect regardless of the actions that are taken.

Our aim here was to adjust the model so that the errors it made were qualitatively similar to those made by human listeners. We therefore focussed on generating Masker and Mix errors. We tested four different hypotheses by which Masker and Mix errors might arise, which are displayed in Figure 3. (1) Imprecise likelihood mappings between *Target Colour* and *Target Number* states and the observed words on the attended side (e.g., *Left Colour Word* or *Right Colour Word* outcomes, depending on the *Spatial Attention* state). From the perspective of model inversion, this mechanism affects the extent to which auditory sensations on the target side update the model’s beliefs about the target words: Lower precisions are associated with less belief updating. (2) Very precise likelihood mappings between *Target Colour* and *Target Number* states and observed words on the *unattended* side. Higher precisions for words on the unattended side allows auditory sensations that are unattended to influence beliefs about target words, and model an inability to attend away from the location that was not cued. (3) Imprecise likelihood mapping between *Spatial Attention* states and the observed *Visual Cue*. From the perspective of model inversion, this mechanism reflects the extent to which seeing a Left visual cue enables the model to direct spatial attention to words on the left rather than the right side: Low precision means that sometimes words on the right side are attended when the spatial cue indicates that the target words are on the left. (4) Imprecise likelihood mapping between *Response* states and the *Feedback* outcome.

We expected imprecise beliefs about the relationship between the *Response* states and *Feedback* outcomes to lead to absent errors, for the reason described above and, thus, Hypothesis 4 was included in the simulations as a comparison. Whereas, we expected that any of Hypotheses 1–3 could be associated with Masker and Mix errors.

We varied each of the ensuing four likelihood precisions parametrically. For each mechanism, we simulated 26 different levels (between 0 and 5, except for words on the unattended side, which varied between 0 and 1024), and simulated 48 trials at each level. Likelihood precisions were specified with a precision parameter that determines the relationship between on-diagonal and off-diagonal elements of the relevant A-matrices in the model: the precision (*p*) is the temperature parameter of a softmax function on an identity matrix (where *N* is the number of state factors):

$$\mathbf{A} = \frac{1}{N-1+e^{p}} \begin{bmatrix} e^{p} & 1 & \cdots & 1 \\ 1 & e^{p} & \cdots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & 1 & e^{p} \end{bmatrix} \qquad (1)$$

To disambiguate the effects of different mechanisms, we varied each precision while keeping the other precisions constant. We used precisions for the other parameters that were either low (in the case of words on the unattended side: its associated precision parameter was set to zero when it was not the factor of interest) or high (for all other factors, precision was set to 1024 when they were not factors of interest) (see Figure 3). We set the value of the alpha parameter to one, so that actions were selected according to the inferred probability of allowable actions.

For each of the above mechanistic hypotheses, we examined effects of precision on the percent of trials that had a correct response and, on incorrect trials, the percentages of each error type (Masker, Mix, and Absent). The results are shown in Figure 4. Chance performance is 6.25%, because there are 16 possible policies and only one is correct.

< FIGURES 3 & 4 >

Lower precision of the likelihood mappings between *Target Colour* and *Target Number* states and observed words on the *attended* side systematically engendered worse performance (Figure 4A: left panel)—particularly when the precision parameter was less than 1. However, the majority of errors were Absent errors (Figure 4A: right panel). When the precision of the likelihood mappings between *Target Colour* and *Target Number* states and observed words on the *unattended* side was greater than 0, performance was low although still above chance (Figure 4B: left panel). Unlike the first manipulation, however, most of the errors were Mix errors, and the second highest error type were

Masker errors (Figure 4B: right panel). The precision of the likelihood mapping between *Spatial Attention* states and the observed *Visual Cue* only seemed to affect performance when the mapping was entirely imprecise (i.e., zero precision, which indicates a uniform matrix), which led to mainly Absent errors (Figure 4C). Finally, changing the precision of the likelihood mapping between *Response* states and the *Feedback* outcome had a more incremental effect: performance was worse for lower precisions and was at chance level when the precision was zero (Figure 4D: left panel). As expected, this manipulation led to predominantly Absent errors (Figure 4D: right panel).

Perhaps surprisingly, the pattern of errors for three of the four manipulations was similar (Figure 4A,C,D), characterised by a high percentage of Absent errors and low percentages of Masker and Mix errors. Only one of the manipulations—increasing the precision of the likelihood mappings between *Target Colour* and *Target Number* states and observed words on the *unattended* side—led to a higher percentage of Mix errors than other errors. This profile is similar to that seen in behavioural studies (Holmes et al., 2018; Johnsrude et al., 2013; Maddox & Shinn-Cunningham, 2012: "Consistent" condition).

These results show how different likelihood precisions can underwrite error types in this simple cocktail party listening paradigm. We found that reducing the likelihood precision for words on the attended side (Hypothesis 1) did not lead to a high proportion of Masker and Mix errors, and instead led to Absent errors. Conceptualising this using cognitive terms, if we reduce attention to words on the attended side, this does not necessarily mean that words from the unattended side will be reported, if the precision for words on the unattended side is low as in these simulations; instead, this leads to reporting words at random. Similarly, reducing the likelihood precision for the visual cue (Hypothesis 3) led to mainly Absent errors—although, interestingly, the simulations were relatively robust to a severe reduction in precision. In other words, even if the left spatial cue is only weakly associated with attending to the left side, the correct words are usually still reported. Interestingly, however, it was the manipulation of precision for words on the unattended side (Hypothesis 2) that led to Masker and Mix errors—suggesting that when the precision for words on the unattended side was sufficiently high, these words could 'break through' to influence the reported words.

While previous empirical studies are consistent in reporting the greatest percentage of Mix errors, they differ in the relative percentages of Marker and Absent errors. Our manipulation of the precision parameter related to words on the unattended side showed a greater percentage of Masker Errors than Absent errors, which differs from the findings reported by Holmes et al. (2018), upon which the current paradigm was based. However, this pattern aligns with the results reported by Johnsrude et al. (2013). Maddox and Shinn-Cunningham (2012) show that the relative percentages of errors depend on whether the configuration of talkers differs within a spoken phrase, and so is expected to differ under different versions of the paradigm. In addition, it is likely that human patterns of errors are not only attributable to a single mechanism. For example, if a large proportion of errors were made because of an overly precise likelihood mapping to words on the unattended side, and a smaller proportion of errors were made due to an imprecise response-to-feedback likelihood mapping, then this could lead to the highest proportion of Mix errors, a smaller proportion of Absent errors, and the smallest proportion of Masker errors. In other words, human errors may result from a combination of processes. Regardless, here we show that a high percentage of Mix errors can only be explained by high precision associated with words on the unattended side, and not by the remaining precision manipulations that we considered.

To examine the combined (interactive) effects of the precision of the A-matrix for words on the attended and unattended sides (i.e., manipulations 1 and 2, above), we ran simulations at combinations of 21 different levels of each factor between 0 and 2. In other words, we manipulated the two parameters in combination rather than in isolation to examine their non-additive effects.

Figure 5 shows that these two precision parameters have an interactive effect on percent correct. When the precision for words on the attended side is much greater than the unattended side (upper left corner of Figure 5A), accuracy is 100%. However, when the precision for words on the attended side approaches the precision for words on the unattended side (i.e., near to the diagonal in Figure 5A), percent correct decreases. This is not all-or-none, but is rather a graded effect that appears when the difference in precision is less than approximately 0.5, suggesting that the magnitude of the difference in precision affects performance. As the precisions converge (i.e., become more similar), this produces mainly Mix errors and some Masker errors (Figure 5B–D)—showing that words from the unattended side are able to 'break through' to influence the agent's beliefs about target words. The exception is when the precision of both mappings is low (i.e., towards the lower left corner of the plots): in this scenario, the majority of errors are Absent errors (Figure 5D).

< FIGURE 5 >

Overall, these simulations demonstrate that the sort of errors exhibited by human listeners (i.e., masker and mix errors) occur when the precision for words on the unattended side is only marginally lower than the precision for words on the attended side. In other words, errors occur when participants are not assigning sufficiently high precision to the attended side (or, conversely, sufficiently low precision to the unattended side) to prevent 'break through' from words on the unattended side. This means that, for these types of errors to occur, attention does not necessarily need to be misallocated (i.e., attending to the non-cued location rather than the cued location)—but, instead, it is simply not allocated 'strongly' enough (i.e., with sufficiently high precision) to suppress breakthrough from words on the non-cued side.

Modelling preparatory attention: Reaction times

In the simulations that follow, our aim was to simulate preparatory attention; specifically, the apparent build-up of attention over time before the onset of a target stimulus. This reflects the more general finding that attention is not all-or-none, and we wanted to examine how this type of non-binary voluntary attention could be modelled. With this goal in mind, we sought to reproduce two sets of empirical findings for preparatory attention that speak to time-varying attention: the systematic shortening of reaction times with longer cue-target intervals (reported by Holmes *et al.,* 2018), and a build-up of ERP responses during the cue-target interval (reported by Holmes et al., 2017). Crucially, both of these findings were obtained using the same paradigm as we described above. In the previous section, precision parameters were fixed over time. Whereas, we predicted that—in order to reproduce behavioural and electrophysiological correlates of preparatory attention—we would need to simulate changes in precision over time. The rationale behind this prediction is that likelihood mappings have more time to become precise when the cue-target interval is longer, and more precise mappings are likely to lead to faster reaction times. Therefore, in this section and the subsequent section, we consider time-varying precision. The neural basis for time-varying precision could be changes in neuronal gain (Parr & Friston, 2019b).

In this section, we focus on the shortening of reaction times with longer cue-target intervals. In active inference, response times are the time taken for the gradient descent on free energy to converge. In more general terms, they can be considered as a measure of statistical efficiency, under the Principle of Least Action (because Action in physics is a time integral). We tested competing hypotheses about how precision evolves over time. We predicted that a faster build-up of precision would translate into faster belief updating, leading to faster convergence within each epoch (i.e., time-step within a trial) and hence faster response times (RTs).

To model different temporal profiles, we simulated a long trial (consisting of 14 time-steps) in which the spatial cue and the spoken words were presented at different times. To model changes in precision, we introduced an additional state factor into the generative model: *Attentional Focus* (see Figure 6). This factor equipped the model with an explicit belief about the precision for words on the attended side, which could, for example, relate to a belief about when the target talker will start speaking.

The A-matrix was specified so that the *Attentional Focus* states determined the likelihood precision between *Target Colour* states and observed colour words (i.e., *Left Colour Word* or *Right Colour Word* outcomes, depending on the *Spatial Attention* state). The same precision value was used for the mappings between *Target Number* states and observed number words. The B-matrix for the *Attentional Focus* state was specified as a shift matrix (i.e., ones on the sub-diagonal and zeros elsewhere) with the additional specification that the final (i.e., most focussed) state is an absorbing state. Consequently, the attentional state transitions through the *Attentional Focus* states, in turn, as the trial proceeds, until the most focussed stated is reached. The agent then remains in the most focussed state until the trial ends. This reflects the implicit prior belief that, during a trial of fixed length, we can be increasingly confident that the stimulus is imminent as time elapses, if we have not yet observed the stimulus.

In these simulations, we separated the generative process (i.e., the process generating outcomes in the world) from the generative model (i.e., the synthetic agent's internal generative model of the task), because attention only operates in the brain and not in the world. Thus, we let *Attentional Focus* modulate precision in the generative model, but not the generative process.

Notice that the *Attentional Focus* and *Spatial Attention* states are factorised in this generative model. That is, the *Attentional Focus* states only encode the precision of the mapping between word states and outcomes, not which words (i.e., left or right side) are attended. This could be thought of as a distinction between 'what' (*Spatial Attention*) and 'when' (*Attentional Focus*)—which interact to produce a particular attentional set (Auksztulewicz et al., 2018; Coull & Nobre, 1998). We revisit this point in the discussion.

We included 10 discrete *Attentional Focus* states, and each state was associated with a particular precision value: we specified distinct precisions in different models to characterise the temporal profiles that could plausibly generate a slow instantiation of attentional set. In other words, for the purposes of these simulations, we discretised time between the presentation of the visual cue and the spoken word into 10 bins. We could have simulated smaller time steps, and thus a more gradual change in *Attentional Focus* over time, but this was unnecessary for the behavioural simulations that follow.

< FIGURE 6 >

We tested 5 different hypotheses about how precision changes over time, which are illustrated in the lower panel of Figure 6: (i) Precision increases linearly; (ii) Precision follows an exponential function; (iii) Precision follows an exponential cumulative density function; (iv–v) Precision is constant over time (i.e., null hypothesis). We can think of hypotheses (i)–(iii) as progressing from an imprecise state to a maximally precise state—which means that greater precision was allocated to words on the attended side as time progressed. These hypotheses differ, however, in the shape of the time-dependent increase in precision. Hypothesis (i) assumes that precision increases steadily from the time at which the cue occurs until 2000 ms after the cue is presented. Hypothesis (ii) assumes that precision begins low, then rapidly increases shortly before 2000 ms. Hypothesis (iii) assumes that precision increases rapidly soon after the cue is presented and remains high. We also specified two alternatives for the null hypothesis: one corresponding to a low uniform precision (hypothesis (iv)), and a second corresponding to a high uniform precision (hypothesis (v)). We chose to include two null hypotheses rather then one, so that the null hypothesis was not biased towards a particular precision. For each hypothesis, we selected 10 different (equally spaced) positions on the temporal function, which were discretised into 10 different *Attentional Focus* states. The corresponding precision values specified the temperature parameter on a softmax (normalised exponential) function, which specified the relative difference between on-diagonal and off-diagonal elements of the A-matrix (see Eq. 1), as in the previous section.

In these simulations, we assumed that the cue was revealed in the first epoch. Essentially, this meant that we did not model the period of time before the cue was presented, which contained no relevant information for the task. To simulate the different lengths of preparation time used in the experiment by Holmes *et al.* (2018), we changed the epoch when the talkers started speaking—essentially simulating different cue-target intervals (i.e., the stimulus-onset asynchrony). Thus, for each preparation time condition, the agent was in a different *Attentional Focus* state at the time that the talkers started speaking. In other words, the precision value when the talkers started speaking was at a different place on the temporal functions illustrated in Figure 6(i)–(v). We assumed that each epoch corresponded to 250 milliseconds, which was the shortest preparation time used in Holmes *et al.* and the greatest common factor of all of their preparation times (250, 500, 1000, and 2000 ms). We simulated 0 ms preparation time by specifying outcomes as words at the start of the trial, on the second epoch (the trial always begins with the agent hearing nothing on the first epoch, as described above). We simulated preparation times of 250, 500, 1000, and 2000 ms by introducing the talker onset at the third, fourth, sixth, and tenth epochs. In this paper, we refer to this variable as the 'Talker Onset Epoch'.

We added a null outcome for each of the word outcomes (see Figure 6), to indicate that the talkers were not speaking any words at the beginning of the trial. We also added a null *Feedback* outcome: When no response was made, feedback was neither correct nor incorrect. The null *Feedback* outcome was considered to be a non-preferred outcome (see Appendix 3 for parameter values), so the synthetic agent should try to change the outcome by making a response, but a null outcome was still preferred over an incorrect outcome.

Given we were interested in reaction times, our policy space covered the different actions the agent could make (one of 16 response buttons) and *when* the agent could make the action (i.e., in which epoch): Thus, the agent could decide to respond early in the trial (e.g., before the talker started speaking) or wait to accumulate evidence before changing from a Null response to one of the 16 colour-number combinations. This meant that we had a large policy space (208 policies: 16 response options at 13 epochs; the agent was not allowed to respond in the first epoch).

During model inversion under active inference, a response time is calculated for each epoch. As our response time measure, we used the integral (i.e., sum) of response times from the epoch at which the talkers started speaking until the epoch that the action changed from Null to one of the 16 colour-number combinations. This measure accounts for both the *number* of epochs until the response was enacted and the efficiency of the gradient descent *within each epoch*.

For each hypothesis, we simulated average RTs, and directly compared these to the experimental data reported by Holmes *et al.* (2018). We did this by calculating the average RT, as described above, for each Talker Onset Epoch. For each Talker Onset Epoch within each of the five hypotheses, we simulated 100 trials. After taking the average over trials, we linearly scaled the simulated RTs—which are in arbitrary units—to the group data reported by Holmes *et al.* (2018), so that the response times at Talker Onset Epochs of 2 and 10 matched the experimental data for preparation times of 0 and 2000 ms, respectively. This allowed us to contrast the shapes of the temporal RT functions among the 5 hypotheses.

The results are shown in Figure 7A, alongside the data from Holmes *et al.* (2018). Contrary to our prediction, the temporal functions are extremely similar among the 5 hypotheses. All of the simulated RTs are within one standard deviation of the experimental data.

Figure 7B shows the root-mean-squared (RMS) error between each of the 5 hypotheses and the experimental data. The RMS error is low for all hypotheses, suggesting that all hypotheses explain the data well, but is marginally lower for the low-precision uniform (i.e., null) hypothesis. It is interesting to note that the two uniform (null) hypotheses are the simplest of the five hypotheses, and do not require the presence of an *Attentional Focus* state. Thus, these simulations demonstrate that time-dependent changes in precision are unnecessary to explain the RT data reported by Holmes *et al.* (2018).

< FIGURE 7 >

This raises the question: What underlies the difference in RTs between the five preparation time conditions (modelled as five Talker Onset Epoch conditions), if it is not mediated by time-sensitive precision (which differed under the five hypotheses)? A possible explanation is that faster RTs reflect the evaluation of a smaller number of policies, because the efficiency of gradient descent is related to the number of plausible policies that are retained. In active inference, policies are pruned if the evidence for the policy falls outside of Occam's window (i.e., its posterior probability is very low). Thus, we subsequently examined whether policy pruning could account for differences in simulated RTs among Talker Onset Epoch conditions.

Figure 8 shows the time-course of policy pruning in an example trial. Figure 8A shows that every Talker Onset Epoch condition began with the same number of policies and finished with one (winning) policy, but many of these were pruned away during the first few epochs. Relative to the onset of the trial, policies were pruned slightly (although not substantially) earlier when the talker onset was earlier. This seems intuitive because the response depends on the spoken words, which are not revealed until the talker onset epoch, and cannot be inferred from the attentional cue alone. To account for the difference in the epoch at which the talker outcomes were revealed among the different conditions, Figure 8B plots policy pruning relative to talker onset. This reveals systematic differences between policy pruning for the five different Talker Onset Epoch conditions. Note that the way of plotting policies in Figure 8B (i.e., relative to the onset of the target stimulus) better corresponds to how RTs are typically calculated in empirical studies, than those displayed in Figure 8A.

< FIGURE 8 >

Table 1. Results of the hierarchical linear regression, showing how the number of policies at each epoch affects the reaction time (RT) for the trial. This analysis incorporates simulated trials under all models. Each row shows a different model, incorporating different variables (which were entered using the stepwise method, meaning that the variables that accounted for the greatest variance were entered first). R^2 change and p-values result from comparing the model in that row with the model in the row above (not applicable for the first row).

Variables in model	R^2 (%)	R^2 change (%)	p-value
Policies in epoch 1	95.3	-	-
Policies in epochs 1 & 2	97.2	1.9	< .001
Policies in epochs 1, 2 & 3	97.3	.1	< .001
Policies in epochs 1, 2, 3 & 5	97.4	.1	< .001
Policies in epochs 1, 2, 3, 4 & 5	97.6	.2	< .001

Generally speaking, when the talker onset epoch is earlier, there are more policies within Occam's window at the talker onset epoch (i.e., at epoch 1 in Figure 8B)—and it takes a greater number of epochs (i.e., more rounds of belief updating) to prune the set of policies to one winning policy. In general, these simulations imply that the RT benefit of lengthening the cue-target interval in the current simulations was that, as time passes, fewer policies are plausible, and so policy selection can occur sooner after talker onset. To test this hypothesis, we entered the number of policies on each simulated trial (at epochs 1–5 relative to talker onset) as predictor variables in a general linear model (GLM), with simulated RTs as the dependent variable. A hierarchical stepwise regression showed that the number of policies at the epoch corresponding to the talker onset (i.e., epoch 1) was a significant predictor of simulated RT (p < .001), accounting for 95.3% of the variance in RTs across all of the simulated trials. Although the number of policies at epoch 1 was the best predictor of RTs, including the number of policies at subsequent epochs significantly improved the model fit further (Table 1).

This relationship was conserved when each of the 5 models were tested in separate hierarchical regressions, although there were some subtle differences in the ordering of variables. For all models, the number of policies in epoch 1 accounted for the most variance in RTs (R^2 = 94.8%–94.8%, p < .001). For the uniform model with lower precision, the number of policies in epoch 4 was entered second into the regression (R^2 change = 2.9%, p < .001), and none of the other variables significantly improved the model fit beyond this (p ≥ .15). For the other 4 models, the number of policies in epoch 2 was entered second (R^2 change = 2.1–2.3%, p < .001), and the number of policies in epoch 5 was entered third (R^2 change = .1–.2%, p < .001). None of the other variables significantly improved the model fit beyond this (p ≥ .06), apart from the exponential CDF model, for which adding the number of policies in epoch 4 improved the model fit beyond the number of policies in epochs 1, 2, and 5 (R^2 change = .02%, p = .031).

Overall, these results imply that the RT benefit at longer preparation times can be accounted for by policy pruning, rather than greater precision. In a speeded reaction time task, an artificial listener begins with a large number of policies, reflecting the fact that they expect to respond at any time from when the cue is presented until the trial ends. However, as the listener waits for the target talker to start speaking, they can discount policies that involve responding during the preparatory interval, because—if they have not yet heard the talker speak—there is insufficient evidence for those policies. When the cue-target interval is longer, the listener has already discounted a greater number of policies by the time that the target talker starts speaking, so the correct policy can be selected more quickly after talker onset. Whereas, when the cue-target interval is short, the artificial listener still has lots of possible policies to evaluate (some for responding quickly and others for responding more slowly), and this slows down the process of belief updating (in the Bayesian model comparison literature, this is sometimes known as model dilution). In our model, the pruning of policies was sufficient to account for different RTs with different lengths of preparation time. We found no improvement in model fit when we included time-dependent changes in precision.

Modelling preparatory attention: EEG responses

Finally, we examined simulated EEG responses under the five temporal hypotheses, to test whether time-dependent changes in precision are necessary to account for EEG responses observed during preparatory attention. In particular, we evaluated the ability of the five models to predict the CNV response (Tecce, 1972; Walter et al., 1964), which has been observed in a variety of domains, in both the auditory and visual modalities (Pasinski, Mcauley, & Snyder, 2016; Walter et al., 1964). The CNV is a negative potential over fronto-central electrodes that gradually increases in amplitude over time, until a target stimulus occurs. It is thought to reflect the anticipation of an upcoming stimulus (Chennu et al., 2013), and larger magnitude CNVs are related to better detection of acoustic stimuli (Rockstroh, Müller, Wagner, Cohen, & Elbert, 1993). CNV-like responses have been observed during preparatory attention for cocktail party listening, as characterised by an increase in amplitude before a target talker begins to speak (Holmes et al., 2016, 2017).

Given the simulated RTs presented in the previous section, we asked whether the five hypotheses predict similar EEG responses, or whether the temporal profile of simulated EEG responses differs among hypotheses—resembling the CNV under some hypotheses more than under others. In the EEG simulations in this section, we focussed on the longest Talker Onset Epoch, which emulates empirical analyses of event-related potentials (ERPs) in preparation for cocktail party listening (Holmes et al., 2016, 2017).

In active inference, EEG responses directly relate to belief updating under the generative model. The simulated EEG response is the rate of change of neuronal activity encoding the sufficient statistics of posterior beliefs (Friston, FitzGerald, et al., 2017). This activity is simulated using a gradient descent on free energy (or, in other words, a gradient ascent on marginal likelihood). We simulated a trial under each of the five models, and extracted the accompanying neuronal dynamics. We then filtered the simulated neuronal responses as if they were an ERP and compared these simulated responses to the empirical ERPs reported in Holmes et al. (2017; Figure 2C: Cluster 2N). For further details about simulated neuronal dynamics under active inference, please see Friston et al. (2017).

Figure 9 shows the simulated neural responses for the five hypotheses. All of the simulated responses show a similar theta component, which simply arises because we scheduled belief updating every 250 ms; in other words, the synthetic agent updates their beliefs about which *Attentional Focus* state they are in every 250 ms. This theta component is therefore of less interest here, although it might be considered to be similar to the time-scale at which people sample sensory data (e.g., rates of syllable presentation or the timing of saccades; Friston et al., 2012; Giraud & Poeppel, 2012). Crucially, the slower drifts in the amplitude of EEG activity (which are illustrated in red in Figure 9, for visualisation purposes) differ among models, consistent with their distinct precision profiles. Although there were no new outcomes before the talkers started speaking (which occurred at 2.25 seconds in Figure 9), the synthetic agent updated its beliefs about the *Attentional Focus* state, which transitioned throughout this time period—in other words, reflecting changes in precision that were associated with different *Attentional Focus* states.

< FIGURE 9 >

The models with uniform precision (Figure 9D–E) give rise to ERPs that have a stable amplitude. The ERPs for the linear model (Figure 9A) show an early increase in amplitude until approximately 1.25 seconds, and then a plateau. The exponential CDF model (Figure 9C) shows an early increase similar to the linear model, but then a sharper decrease in amplitude towards the baseline amplitude. The exponential model (Figure 9B) shows a delayed increase in amplitude that peaks around the time of talker onset. Based on visual inspection of Figure 9, the exponential model is most consistent with the ramping of ERP amplitude during preparatory attention, of the sort attributed to the CNV response.

To formally compare the models, we entered the simulated ERPs from each of the 5 models into a hierarchical linear regression (stepwise method) to predict the data reported in Holmes et al. (2017). We selected Cluster 2N from Holmes et al. (2017), because this was the cluster that displayed the CNV-like response. It is worth noting that their analysis identified a complementary cluster (Cluster 2P) with opposite (positive) polarity, which simply reflects differences in polarity when ERPs are measured at different places on the scalp relative to the source of activity. We extracted the empirical data for Cluster 2N between 0 and 2000 ms, which corresponds to the time period of the cue-target interval in Holmes et al. (2017). We compared these data with our simulated ERPs between 0.25 and 2.25 seconds.

We found that the exponential CDF model explained the most variance in the empirical EEG data ($R^2 = .42$, $\beta = .53$, $p < .001$), and the exponential model explained additional variance unaccounted for by the exponential CDF model (R^2 change = .08, $\beta = -.31$, $p < .001$). None of the responses from the remaining models accounted for additional variance ($p > .38$). The exponential CDF model appears to account best for the early response to the visual cue, measured within the first 1000 ms after the cue is presented, whereas the exponential model accounts best for the later CNV-like response at 1000–2000 ms, which is immediately before the target talker starts to speak.

Overall, these simulated data show that, while time-dependent changes in precision are unnecessary to explain the RT advantage at longer preparatory intervals (as reported in the previous section), they are needed to explain the ramping pattern of EEG activity—suggesting that, in fact, these two empirical findings are not necessarily underpinned by the same computational process. Of the models we tested, an exponential increase in precision best approximates the empirical CNV response. This implies a time-dependent precision in which precision begins low, then rapidly increases shortly before the target talker starts speaking.

Discussion

In this paper, we introduce a generative model that accounts for spatial attention during a simplified cocktail party paradigm that uses CRM phrases. Active inference under this generative model offers 100% accuracy in a spatial cueing task, but performance is impaired in different ways when the precisions of various likelihood mappings in the generative model are changed, revealing computational dissociations. Crucially, we found that masker and mix errors during cocktail party listening can be accounted for by a likelihood mapping for words on the attended side that is only marginally more precise than the likelihood mapping for words on the unattended side. This allows words on the unattended side to ‘break through’ to update beliefs about target words. Thus, a low precision for attended relative to unattended words might explain why human listeners typically make masker and mix errors during cocktail party listening, rather than reporting other words that are not in the mixture (Best et al., 2008; Brungart & Simpson, 2002; Holmes et al., 2018). This speaks to an explanation of masker and mix errors, whereby attention does not need to be misallocated to the non-target location, but may simply not be allocated strongly enough to the attended location relative to the unattended location. The precision of likelihood mappings is generally understood as neuronal gain (Parr & Friston, 2019b); thus, the difference in precision could conceivably be reflected in the gain of neuronal populations encoding words on the unattended side, relative to words on the unattended side. We found that Absent errors could be accounted for by a variety of 'computational lesions' under our generative model, reflecting the variety of processes that can lead participants to report words that were not spoken on a trial. For example, a lack of motivation to respond correctly, or failures at the response stage. Given that a variety of errors are often observed in auditory attention experiments, future work could employ this generative model as a tool for dissecting the processes underlying these errors, and how they vary across participants: Equipped with each participants' behavioural responses, this generative model could be inverted to recover the most likely underlying parameters that account for participants' responses and, ultimately, which parameters best account for individual differences or differences between groups of participants (Schwartenbeck and Friston, 2016).

When we extended the generative model to accommodate preparatory attention, we found that time-dependent changes in precision were not needed to explain faster RTs with longer preparatory intervals, but were necessary to explain the ramping of ERPs (resembling the CNV) before the target talker begins to speak. We found that faster RTs could be explained quite simply by the pruning (i.e., discounting) of policies over time—and time-dependent changes in precision did not improve the fit of our simulations to empirical RTs from human listeners. This speaks to an explanation of preparatory RTs that involves a relatively automatic process, related to the need to act quickly (i.e., a psychological Principle of Least Action), rather than to a volitional control of precision that changes over time. For example, it might relate to the conditional probability that the target talker will start speaking, which—given that this has not yet occurred—increases over time (c.f., a harzard function) (Luce, 1986). Alternatively, given that our policy space included both the nature and the timing of the response, it could relate to a speed-accuracy trade-off. The latter would imply that such a pattern of RTs would only be observed in challenging task contexts, where the need to respond quickly is in conflict with the need to respond accurately. Consistent with this idea, visual cues do not improve performance on a cocktail party listening task when the talkers are easily separable based on their acoustic characteristics (Varghese, Ozmeral, Best, & Shinn-Cunningham, 2012).

We found that models with different time-dependent precision profiles give rise to different electrophysiological responses. Our simulated electrophysiological responses constitute updates to the agent's beliefs about the *Attentional Focus* state, which—in our generative model—determines the precision of likelihood mapping for words on the attended side. As shown in Figure 9, flat precision profiles were associated with flat responses, whereas models with time-dependent increases in precision generated increased neuronal responses during the preparatory interval (i.e., before the target talker started speaking). We could clearly demonstrate that different patterns of precision are associated with different amplitude profiles. This is intuitive from the perspective of active inference, because belief updating is related to the rate of change in precision. Of the models we tested, the exponential model best accounted for CNV-like responses that have been observed in EEG studies of preparatory attention. Although we only tested three formally distinct temporal changes in precision here, our findings speak to a temporal deployment of attention that increases non-linearly—beginning low when an attentional cue is presented and rapidly increasing shortly before the target stimulus occurs. This fluctuation in precision is specified as an explicit belief in our generative model, and it could, therefore, be considered as a volitional deployment of attention—which could be instantiated as an increase in neuronal gain—possibly reflecting the time at which a listener expects a target stimulus to occur. Indeed, studies that have measured the CNV using two or more temporal intervals found that the timing of the CNV aligns well with participants' expectations about when a stimulus will occur (Birbaumer, Elbert, Canavan, & Rockstroh, 1990;

Rockstroh et al., 1993; Ruchkin, McCalley, & Glaser, 1977). Also, the CNV has a steeper slope when shorter intervals are used that can be more accurately estimated (McAdam, Knott, & Rebert, 1969; Miniussi, Wilding, Coull, & Nobre, 1999). Although, in these previous studies of the CNV, the CNV is closely associated with a 'preparatory set' related to motor preparedness (Rohrbaugh, Syndulko, & Lindsley, 1976); whereas, in the current cocktail party setting, the cue does not inform participants about the motor response, because the response depends upon the words spoken by the target talker.

While we simulated a simplified cocktail party setting here, similar computational processes might also underpin preparatory and selective attention in other domains. The *Attentional Focus* factor in our generative model could be considered more generally as a possible mechanism for temporal attention (e.g., Shen & Alain, 2011, 2016); specifically, 'temporal orienting' (Coull & Nobre, 1998; Nobre, 2001; Nobre, Correa, & Coull, 2007). Implicit temporal preparation has been shown to affect behaviour even when it is not task-relevant (Vallesi, 2010). In this context, the temporal functions in Figure 6(i)-(v) reflect different expectations about *when* a target stimulus is expected to occur. One could imagine a family of functions, reflecting the time that participants expect an event to occur (and uncertainty about the expected time). Under this formulation, the particular form of time-dependent precision constitutes a hypothesis in the model, and the null hypotheses considered in this paper correspond to a belief that a target is no more likely to occur at one time point than another. Interestingly, the empirical EEG responses that we modelled were recorded in an experiment in which the target always occurred 2000 ms after the cue. Thus, participants would have strong expectations that the target would occur around 2000 ms, which could explain why the exponential temporal function, which increased rapidly around 2000 ms, was the best fit to the CNV-like response. Future experiments could test this more directly by controlling the distribution of cue-target intervals, factorising different expected values and precisions over the length of the cue-target interval, to observe how this affects the temporal profile of electrophysiological responses.

Our model factorises two different types of attentional states. The *Spatial Attention* factor, which is akin to the type of attentional state that has been modelled in previous work (e.g., Feldman & Friston, 2010; Mirza et al., 2019), determined *which* spatial location was the focus of attention. Whereas, the *Attentional Focus* state played a different role: it determined *how strongly* attention was focussed on a particular location, irrespective of which location was attended. In the current simulations, we used the *Attentional Focus* states to control gradations in attention within a trial that were not all-or-none. In this way, the *Attentional Focus* state only modified the strength (i.e., precision) of connections between particular states and outcomes. We could envisage a different model structure, whereby a single factor determines both the strength *and* the direction of attention. The factorisation used here would be more efficient, because many more states are needed to represent the combination of strength and direction (2 directions * 10 strengths = 20 states) than the factorised version (2 directions + 10 strengths = 12 states). When considering the *Attentional Focus* state in the context of temporal attention, one could consider this formulation as a factorisation of spatial and temporal attention—in other words, distinct beliefs about attention to *what* and *when*. This is consistent with evidence from human electrocorticography (Auksztulewicz et al., 2018) showing that predictions about the timing and content of a spoken syllable evokes activity in distinct bran areas, and is best modelled by different changes in synaptic gain. The current model could be extended to incorporate other forms of attention, such as frequency-specific attention, which could be encoded in a similar way as the spatial attention factor.

We assume that the slow drifts in attention that we modelled here—which occur on the order of seconds—are distinct from faster changes in attention that have previously been associated with theta oscillations (Morillon & Schroeder, 2015; Pefkou, Arnal, Fontolan, & Giraud, 2017; Senoussi, Moreland, Busch, & Dugué, 2019; VanRullen, 2018). These faster attentional processes may be associated with the temporal sampling of outcomes—for example, at the rate of syllables in a spoken sentence. Whereas, the temporal changes in attention we modelled here operate over a relatively longer temporal scale (although not as long as sustained attention, which may operate over tens of seconds and is accompanied by even slower changes in neuronal dynamics). One could imagine a hierarchical nesting of temporal scales (Friston, Parr, & de Vries, 2017; Friston et al., 2020; Giraud & Poeppel, 2012; Hovsepyan, Olasagasti, & Giraud, 2020; Poeppel, Idsardi, & Van Wassenhove, 2008)—in which slower changes in attention (e.g., those that we modelled here) operate at a higher level of the model, and faster attentional processes (related to sampling of stimuli) occur at a lower level.

In future work, the current model could be combined with models mapping the acoustic signal to words (Friston et al., 2021) and with models mapping words to semantics (Friston et al., 2020); this could be used to simulate cocktail party paradigms in which semantically meaningful passages are heard, or conversations in which a listener volitionally directs attention to a conversational partner and then switches their attention to a different talker to gather new semantic information. For example, if the other sentence helps to resolve semantic ambiguity or contains salient words, such as the listener's name. While these more naturalistic instantiations of cocktail party listening require additional processes beyond those studied here, we assume that the top-down attention component of cocktail party listening—which we isolate using simple CRM phrases that lack semantic context—also applies to these naturalistic settings. This type of hierarchical model could also be used to investigate interactions between bottom-up and top-down processes during attentional selection, which has previously been modelled for non-speech sounds (Golob, Venable, Anderson, Benzell, & Scheuerman, 2016). A previous model by Golob et al. (2016) incorporated a spatial attention gradient, which could also be incorporated into the current model—and may be particularly useful for modelling talkers who are not as clearly separated in location as in the current simulations.

The ERPs simulated in Figure 9 are also interesting from the perspective of hearing loss. Holmes et al. (2017) showed that the CNV-like response during the cue-target interval is significantly weaker in people with hearing loss than in people with normal hearing, even when people with hearing loss use hearing aids. During the cue-target interval, their ERPs resemble the flat ERP that is associated with uniform precision, rather than the CNV-like responses associated with an exponential increase in precision. One of the most common problems reported by people with hearing loss—including those who use hearing aids—is difficulty listening in noisy places. Combined with the data reported by Holmes et al. (2017), one possible explanation is that they are unable to increase the precision of mappings to attended relative to unattended locations to the same extent as do people with normal hearing. In future work, we plan to use the current generative model to compare computational parameters between people with and without hearing loss. For example, this type of model can enable us to dissociate the hypotheses that people with hearing loss have a different sort of generative model (i.e., with different parameters), from the hypothesis that they use the same generative model, but it is based on different auditory outcomes, which are distorted due to hearing loss.

The model could also be used to simulate a variety of other empirical effects. For example, the cost of switching attention—which is characterised by slower RTs when a listener is instructed to attend to a different stimulus attribute (Nolden, Ibrahim, & Koch, 2018; Seibold, Nolden, Oberem, Fels, & Koch, 2018) or stream (Larson & Lee, 2013; Maddox & Shinn-Cunningham, 2012) than when these remain the same—might be underpinned by a belief that the stimulus set will be enduring. In our Markov decision process, this would be reflected in the B-matrix for the *Spatial Attention* factor. One could equip the model with beliefs that the *Spatial Attention* state will remain the same over time, or beliefs that it will change; we expect that switching to a different attentional state would evoke a greater change in beliefs when the state is expected to remain the same over time, which could slow down subsequent inference and explain slower RTs. The model could also be used to examine object-based attention (Shinn-Cunningham, 2008) by contrasting models in which different features (e.g., location and frequency) are attended separately or in combination. It would also be interesting to examine learning under this generative model: moment-by-moment changes in precision might help to explain why attentional streaming builds up over time (Bregman, 1978). To test any of these hypotheses, Bayesian model comparison could be used to establish which different instantiations of the generative model best account for empirical data (Schwartenbeck and Friston, 2016).

In summary, this paper introduces a generative model that can reproduce empirical behaviour and electrophysiological responses during cocktail party listening, under active inference. We show that this model is useful for predicting effects of different computational deficits on error types, and that errors that involve reporting words spoken by a competing talker—which are a hallmark of informational masking—arise when the likelihood precision for words at an unattended location is only marginally lower than precision for words at an attended location. In addition, we used this model to simulate preparatory attention during the cue-target interval, before a target talker starts to speak. This revealed a dissociation in the computational processes underlying faster RTs with longer cue-target intervals compared with CNV-like ERPs measured during the cue-target interval. The former can be explained by policy pruning and speaks to a relatively automatic process related to the need to act quickly (i.e., a psychological Principle of Least Action). Whereas, the latter can be explained by a temporal evolution of precision in the absence of new sensory observations, which speaks to a volitional deployment of time-sensitive attention that peaks around the time a target talker is anticipated. We envisage that this generative model may be useful for future work examining different aspects of attention and everyday listening, and for examining differences between (groups of) participants.

Author Contributions

EH, TP, TDG, and KJF conceptualised the study. EH conducted the modelling work. EH wrote the first draft of the manuscript. TP, TDG, and KJF edited the manuscript.

Conflict of interest statement

The authors declare no competing interests.

Acknowledgements

This work was supported by a Pauline Ashley Fellowship from RNID to Emma Holmes (PA25_Holmes). This work was conducted at the Wellcome Centre for Human Neuroimaging, which is funded by Wellcome (Ref: 203147/Z/16/Z). Timothy D. Griffiths was supported by WT091681MA and DC000242-31. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We would like to thank Sherry Dongqi Bao for conducting some preliminary modelling work on this paradigm as part of a research visit to UCL.

Appendices

Appendix 1: Free energy equations

Variational free energy (F) is an upper bound on surprise:

$$F = -\ln P(\tilde{o}) + D_{KL}[Q(\tilde{s}) \,||\, P(\tilde{s} \,|\, \tilde{o})] \tag{A.1}$$

where represents the states of the generative model, $\tilde{o} = (o_1, ..., o_\tau)$ denotes sensory observations until the current time (τ), P indicates the probability under the generative model, and Q indicates posterior beliefs. Under this formulation, surprise ($-\ln P(\tilde{o})$) is the negative log probability of observations under the generative model, and the second part of the equation is the Kullback–Leibler divergence (D_{KL}) between the approximate posterior distribution and the true posterior distribution. Thus, minimising variational free energy minimises surprise.

The equation for variational free energy can be written another way as the difference between accuracy and complexity. Thus, minimising variational free energy simultaneously maximises accuracy and minimises complexity.

$$F = \underbrace{D_{KL}[Q(\tilde{s}) \,||\, P(\tilde{s})]}_{complexity} - \underbrace{E_Q[\ln P(\tilde{o} \,|\, \tilde{s})]}_{accuracy} \tag{A.2}$$

where E_Q indicates the expected value under the posterior (Q) distribution.

The enactive component of active inference comes from selecting policies (π; i.e., a series of actions) that minimise expected free energy (G) under the generative model:

$$\begin{aligned} G(\pi) &= \sum_\tau G(\pi, \tau) \\ G(\pi, \tau) &= E_Q[\ln Q(s_\tau \,|\, \pi) - \ln Q(s_\tau \,|\, o_\tau, \pi) - \ln P(o_\tau \,|\, \mathbf{C})] \\ &= \underbrace{D_{KL}[Q(o_\tau \,|\, \pi) \,||\, P(o_\tau \,|\, \mathbf{C})]}_{risk} - \underbrace{H[Q(o_\tau \,|\, s_\tau)]}_{ambiguity} \end{aligned} \tag{A.3}$$

where $\mathbf{C}$ indicates preferences over outcomes, which are documented in Appendix 2.

Minimising expected free energy relies on predictions about future outcomes. The probability of pursuing any particular outcome is proportional to the expected free energy if that action is taken. Expected free energy simultaneously minimises risk (i.e., the Kullback–Leibler divergence between predicted and prior preferences over outcomes under policies) and minimises ambiguity (i.e., the uncertainty about outcomes given states).

For technical details on variational and expected free energy, see Friston et al. (2017).

Appendix 2: Form of the generative model

All of the generative models in this paper take the same generic form:

$$
\begin{aligned}
&P(\tilde{o},\tilde{s},\pi) = P(s_1)P(\pi)\prod_{\tau} P(s_{\tau+1} \mid s_\tau,\pi)P(o \mid s_\tau) \\
&P(o_\tau \mid s_\tau) = Cat(\mathbf{A}) \\
&P(s_\tau \mid s_{\tau-1},\pi) = Cat(\mathbf{B}) \\
&P(o_\tau) = Cat(\mathbf{C}) \\
&P(s_1) = Cat(\mathbf{D}) \\
&P(\pi) = \sigma(-G)
\end{aligned}
\qquad \text{(A.4)}
$$

where $\tilde{s}$ represents the states of the generative model, $\tilde{o} = (o_1,...,o_\tau)$ denotes sensory observations until the current time (τ), π denotes policies, and P indicates the probability under the generative model. The notation Cat means a categorical distribution. Probability distributions are defined by: A-matrices ($\mathbf{A}$), which specify likelihood mappings between states and outcomes; B-matrices ($\mathbf{B}$), which specify how states within each state factor are expected to transition over time; C-matrices ($\mathbf{C}$), which specify preferences over outcomes for each outcome factor; D-matrices ($\mathbf{D}$), which specify prior beliefs about which states the model is likely to be in, initially. The probability of pursuing policies is defined by a softmax (normalized exponential) function (σ) over the expected free energy (*G*).

Appendix 3: Parameters of generative models

Table 2. Relevant parameters of the generative model used in the section "*A generative model of selective attention during cocktail party listening*". Other parameters are specified in the main text or were otherwise set to their default values in SPM12. In the column headed Parameter, colons are used to index a matrix or cell array: they indicate a vector containing every integer between the two integers (inclusive).

Parameter	Value	Description
$\mathbf{C}\{6\}$	[2; -4]	Log preferences over *Feedback* outcomes (respectively: Correct, Incorrect). These values are repeated for all time points in a 2 x T matrix.
$\mathbf{C}\{1{:}5\}$	[1; 1; …]	Log preferences over all other outcome factors: *N*-element vector, where *N* indicates the number of outcomes for each outcome factor. These values are repeated for all time points in a N x T matrix.
α	512	Precision over policies

Table 3. Relevant parameters common to all generative models used in the section "*Modelling impaired performance*". Other parameters are specified in the main text or were otherwise set to their default values in SPM12.

Parameter	Value	Description
$\mathbf{C}\{6\}$	[2; -4]	Log preferences over *Feedback* outcomes (respectively: Correct, Incorrect). These values are repeated for all time points in a 2 x T matrix.

C{1:5}	[1; 1; …]	Log preferences over all other outcome factors: *N*-element vector, where *N* indicates the number of outcomes for each outcome factor. These values are repeated for all time points in a N x T matrix.
α	1	Precision over policies

Table 4. Relevant parameters common to all generative models used in the sections "*Modelling preparatory attention: Reaction times"* and *"Modelling preparatory attention: EEG responses"*. Other parameters are specified in the main text or were otherwise set to their default values in SPM12.

Parameter	Value	Description
C{6}	[2; -4; -3]	Log preferences over *Feedback* outcomes (respectively: Correct, Incorrect, Null). These values are repeated for all time points in a 3 x T matrix.
C{1:5}	[1; 1; …]	Log preferences over all other outcome factors: *N*-element vector, where *N* indicates the number of outcomes for each outcome factor. These values are repeated for all time points in a N x T matrix.
α	512	Precision over policies

References

Auksztulewicz, R., Schwiedrzik, C. M., Thesen, T., Doyle, W., Devinsky, O., Nobre, A. C., … Melloni, L. (2018). Not all predictions are equal: 'What' and 'When' predictions modulate activity in auditory cortex through different mechanisms. *The Journal of Neuroscience*, *38*(40), 0369–18. https://doi.org/10.1523/JNEUROSCI.0369-18.2018

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(35), 13174–13178. https://doi.org/10.1073/pnas.0803718105

Birbaumer, N., Elbert, T., Canavan, A. G. M., & Rockstroh, B. (1990). Slow potentials of the cerebral cortex and behavior. *Physiological Reviews*, Vol. 70, pp. 1–41. https://doi.org/10.1152/physrev.1990.70.1.1

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L., & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: Common computational properties. *Trends in Cognitive Sciences*, *19*(3), 142–150. https://doi.org/10.1016/j.tics.2014.12.008

Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 380–387. https://doi.org/10.1037/0096-1523.4.3.380

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14*(4), 411–427. https://doi.org/10.1007/s10339-013-0571-3

Brown, H., Friston, K. J., & Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in Psychology*, *2*(SEP), 1–10. https://doi.org/10.3389/fpsyg.2011.00218

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101. https://doi.org/10.1121/1.1345696

Brungart, D. S., & Simpson, B. D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *The Journal of the Acoustical Society of America*, *112*(6), 2985. https://doi.org/10.1121/1.1512703

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Chen, Z., Luo, Y., & Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2*(1), 246–250. https://doi.org/10.1109/ICASSP.2017.7952155

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., … Bekinschtein, T. a. (2013). Expectation and attention in hierarchical auditory prediction. *The Journal of Neuroscience*, *33*(27), 11194–11205. https://doi.org/10.1523/JNEUROSCI.0114-13.2013

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustic Society of America*, *25*(5), 1262–2527. https://doi.org/10.1121/1.1408946

Coull, J. T., & Nobre, A. C. (1998). Where and when to pay attention: The neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *The Journal of Neuroscience*, *18*(18), 7426–7435. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9736662

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, *92*, 53–78. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11802865

Feldman, H., & Friston, K. J. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience*, *4*(December), 1–23. https://doi.org/10.3389/fnhum.2010.00215

Fiebelkorn, I. C., & Kastner, S. (2019). A Rhythmic Theory of Attention. *Trends in Cognitive Sciences*, *23*(2), 87–101. https://doi.org/10.1016/j.tics.2018.11.009

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352. https://doi.org/10.1016/j.neunet.2003.06.005

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. J., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*(MAY), 1–20. https://doi.org/10.3389/fpsyg.2012.00151

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation*, *29*, 1–49. https://doi.org/10.1162/NECO_a_00912

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology Paris*, *100*(1–3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network Neuroscience*, 1–78. https://doi.org/10.1162/NETN_a_00018

Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., & Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neuroscience and Biobehavioral Reviews*, *118*, 42–64. https://doi.org/10.1016/j.neubiorev.2020.07.005

Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., & Holmes, E. (2021). Active

Listening. *Hearing Research*, *399*, 107998. https://doi.org/10.1016/j.heares.2020.107998

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Golob, E. J., Venable, K. B., Anderson, M. T., Benzell, J. A., & Scheuerman, J. (2016). Modelling auditory spatial attention with constraints. *International Workshop on Artificial Intelligence and Cognition*.

Helfrich, R. F., Fiebelkorn, I. C., Szczepanski, S. M., Lin, J. J., Parvizi, J., Knight, R. T., & Kastner, S. (2018). Neural Mechanisms of Sustained Attention Are Rhythmic. *Neuron*, *99*(4), 854-865.e5. https://doi.org/10.1016/j.neuron.2018.07.032

Hill, K. T., & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, *20*(3), 583–590. https://doi.org/10.1093/cercor/bhp124

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*(APR), 1–14. https://doi.org/10.3389/fpsyg.2012.00096

Holmes, E., Kitterick, P. T., & Summerfield, A. Q. (2016). EEG activity evoked in preparation for multi-talker listening by adults and children. *Hearing Research*, *336*, 83–100. https://doi.org/10.1016/j.heares.2016.04.007

Holmes, E., Kitterick, P. T., & Summerfield, A. Q. (2017). Peripheral hearing loss reduces the ability of children to direct selective attention during multi-talker listening. *Hearing Research*, *350*, 160–172. https://doi.org/10.1016/j.heares.2017.05.005

Holmes, E., Kitterick, P. T., & Summerfield, A. Q. (2018). Cueing listeners to attend to a target talker progressively improves word report as the duration of the cue-target interval lengthens to 2,000 ms. *Attention, Perception, & Psychophysics*, *80*(6), 1520–1538. https://doi.org/10.3758/s13414-018-1531-x

Hovsepyan, S., Olasagasti, I., & Giraud, A.-L. (2020). Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-16956-5

Ihlefeld, A., & Shinn-Cunningham, B. G. (2008). Disentangling the effects of spatial cues on selection and formation of auditory objects, *J. Acoust. Soc. Am., 124*, 2224–35. https://doi.org/10.1121/1.2973185

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10230795

Kotz, S. A., Schwartze, M., & Schmidt-kassow, M. (2009). Non-motor basal ganglia functions: A review and proposal for a model of sensory predictability in auditory language perception. *Cortex*, *45*(8), 982–990. https://doi.org/10.1016/j.cortex.2009.02.010

Lakatos, P., Barczak, A., Neymotin, S. A., McGinnis, T., Ross, D., Javitt, D. C., & O'Connell, M. N. (2016). Global dynamics of selective attention and its lapses in primary auditory cortex. *Nature Neuroscience*, *19*(12). https://doi.org/10.1038/nn.4386

Larson, E., & Lee, A. K. C. (2013). Influence of preparation time and pitch separation in switching of auditory attention between streams. *The Journal of the Acoustical Society of America*, *134*(2), EL165-71. https://doi.org/10.1121/1.4812439

Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H. M., Larson, E., Hämäläinen, M. S., & Shinn-Cunningham, B. G. (2013). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience*, *6*, 1–9. https://doi.org/10.3389/fnins.2012.00190

Lu, Z.-L., Tse, H. C.-H., Dosher, B. A., Lesmes, L. A., Posner, C., & Chu, W. (2009). Intra- and cross-modal cuing of spatial attention: Time courses and mechanisms. *Vision Research*, *49*(10), 1081–1096. https://doi.org/10.1016/j.visres.2008.05.021

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization. In *Response Times: Their Role in Inferring Elementary Mental Organization*. https://doi.org/10.1093/acprof:oso/9780195070019.001.0001

Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *Journal of the Association for Research in Otolaryngology*, *13*(1), 119–129. https://doi.org/10.1007/s10162-011-0299-7

McAdam, D. W., Knott, J. R., & Rebert, C. S. (1969). Cortical slow potential changes in man related to interstimulus intevval and to pre-trial prediction of interstimulus interval. *Psychophysiology*, *5*(4), 349–358. https://doi.org/10.1111/j.1469-8986.1969.tb02833.x

Miniussi, C., Wilding, E. L., Coull, J. T., & Nobre, A. C. (1999). Orienting attention in time. Modulation of brain potentials. *Brain*, *122*(8), 1507–1518. https://doi.org/10.1093/brain/122.8.1507

Mirza, M. B., Adams, R. A., Friston, K. J., & Parr, T. (2019). Introducing a Bayesian model of selective attention based on active inference. *Scientific Reports*, *9*(1), 1–22. https://doi.org/10.1038/s41598-019-50138-8

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene Construction, Visual Foraging, and Active Inference. *Frontiers in Computational Neuroscience*, *10*(56). https://doi.org/10.3389/fncom.2016.00056

Moore, T. J. (1981). Voice communication jamming research. *AGARD Conference Proceedings 331: Aural Communication in Aviation*, 2:1-2:6. Neuilly-Sur-Seine, France.

Morillon, B., & Schroeder, C. E. (2015). Neuronal oscillations as a mechanistic substrate of auditory temporal prediction. *Annals of the New York Academy of Sciences*, *1337*(1), 26–31. https://doi.org/10.1111/nyas.12629

Necker, L. A. (1832). Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *1*(5), 329–337. https://doi.org/10.1080/14786443208647909

Nobre, A. C. (2001). Orienting attention to instants in time. *Neuropsychologia*, *39*(12), 1317–1328. https://doi.org/10.1016/S0028-3932(01)00120-8

Nobre, A. C., Correa, A., & Coull, J. (2007, August 1). The hazards of time. *Current Opinion in Neurobiology*, Vol. 17, pp. 465–470. https://doi.org/10.1016/j.conb.2007.07.006

Nolden, S., Ibrahim, C. N., & Koch, I. (2018). Cognitive control in the cocktail party: Preparing selective attention to dichotically presented voices supports distractor suppression. *Attention, Perception, and Psychophysics*, 727–737. https://doi.org/10.3758/s13414-018-1620-x

Parr, T., Corcoran, A. W., Friston, K. J., & Hohwy, J. (2019). Perceptual awareness and active inference. *Neuroscience of Consciousness*, *2019*(1). https://doi.org/10.1093/NC/NIZ012

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, 7(1), 1–21. https://doi.org/10.1038/s41598-017-15249-0

Parr, T., & Friston, K. J. (2019a). The computational pharmacology of oculomotion. *Psychopharmacology*. https://doi.org/10.1007/s00213-019-05240-0

Parr, T., & Friston, K. J. (2019b, October 1). Attention or salience? *Current Opinion in Psychology*, Vol. 29, pp. 1–5. https://doi.org/10.1016/j.copsyc.2018.10.006

Pasinski, A. C., Mcauley, J. D., & Snyder, J. S. (2016). How modality specific is processing of auditory and visual rhythms? *Psychophysiology*, *53*(2), 198–208. https://doi.org/10.1111/psyp.12559

Pefkou, M., Arnal, L. H., Fontolan, L., & Giraud, A.-L. (2017). θ-Band and β-Band Neural Activity Reflects Independent Syllable Tracking and Comprehension of Time-Compressed Speech. *The Journal of Neuroscience*, *37*(33), 7930–7938. https://doi.org/10.1523/JNEUROSCI.2882-16.2017

Poeppel, D., Idsardi, W. J., & Van Wassenhove, V. (2008, March 12). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 363, pp. 1071–1086. https://doi.org/10.1098/rstb.2007.2160

Poljac, E., & Yeung, N. (2012). Dissociable Neural Correlates of Intention and Action Preparation in Voluntary Task Switching. *Cerebral Cortex*, (1986). https://doi.org/10.1093/cercor/bhs326

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance X: Control of Language Processes*, *32*, 531–556.

Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the Free-Energy Principle a Formal Theory of Semantics? From Variational Density Dynamics to Neural and Phenotypic Representations, *Entropy, 22*, 889. https://doi.org/10.3390/e22080889

Rockstroh, B., Müller, M., Wagner, M., Cohen, R., & Elbert, T. (1993). "Probing" the nature of the CNV. *Electroencephalography and Clinical Neurophysiology*, *87*(4), 235–241. https://doi.org/10.1016/0013-4694(93)90023-O

Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1976). Brain wave components of the contingent negative variation in humans. *Science*, *191*(4231), 1055–1057. https://doi.org/10.1126/science.1251217

Rolke, B., & Hofmann, P. (2007). Temporal uncertainty degrades perceptual processing. *Psychonomic Bulletin and Review*, *14*(3), 522–526. https://doi.org/10.3758/BF03194101

Rubin, E. (1915). *Synsoplevede Figurer*.

Ruchkin, D. S., McCalley, M. G., & Glaser, E. M. (1977). Event related potentials and time estimation. *Psychophysiology*, *14*(5), 451–455. https://doi.org/10.1111/j.1469-8986.1977.tb01311.x

Schwartenbeck, P., & Friston, K. (2016). Computational phenotyping in psychiatry: A worked example. *eNeuro, 3*(47). doi:10.1523/ENEURO.0049-16.2016

Seibold, J. C., Nolden, S., Oberem, J., Fels, J., & Koch, I. (2018). Intentional preparation of auditory attention-switches: Explicit cueing and sequential switch-predictability. *The Quarterly Journal of Experimental Psychology*, *71*(6), 1382–1395. https://doi.org/10.1080/17470218.2017.1344867

Senoussi, M., Moreland, J. C., Busch, N. A., & Dugué, L. (2019). Attention explores space periodically at the theta frequency. *Journal of Vision*, *19*(5), 1–17. https://doi.org/10.1167/19.5.22

Shen, D., & Alain, C. (2011). Temporal attention facilitates short-term consolidation during a rapid serial auditory presentation task. *Experimental brain research, 215*(3), 285-292. https://doi.org/10.1007/s00221-011-2897-3

Shen, D., Ross, B., & Alain, C. (2016). Temporal cuing modulates alpha oscillations during auditory attentional blink. *European Journal of Neuroscience, 44*(2), 1833–1845. https://doi.org/10.1111/ejn.13266

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Szabó, B. T., Denham, S. L., & Winkler, I. (2016). Computational models of auditory scene analysis: A review. *Frontiers in Neuroscience*, *10*(NOV), 1–16. https://doi.org/10.3389/fnins.2016.00524

Tecce, J. (1972). Contingent negative variation (CNV) and psychological processes in man. *Psychol. Bull.*, *77*(2), 73–108.

Vallesi, A. (2010). Neuro-anatomical substrates of foreperiod effects. In Anna C Nobre & J. T. Coull (Eds.), *Attention and Time* (pp. 303–316). https://doi.org/10.1093/acprof:oso/9780199563456.001.0001

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*.

VanRullen, R. (2018). Attention Cycles. *Neuron*, *99*(4), 632–634. https://doi.org/10.1016/j.neuron.2018.08.006

Varghese, L. A., Ozmeral, E. J., Best, V., & Shinn-Cunningham, B. G. (2012). How visual cues for when to listen aid selective auditory attention. *Journal of the Association for Research in Otolaryngology*, *13*(3), 359–368. https://doi.org/10.1007/s10162-012-0314-7

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). Contingent negative variation: An electric sign of sensori-motor association and expectancy in the human brain. *Nature*, *203*(4943), 380–384. https://doi.org/10.1038/203380a0

Yamaguchi, S., Tsuchiya, H., & Kobayashi, S. (1994). Electroencephalographic activity associated with shifts of visuospatial attention. *Brain*, *117*, 553–562. https://doi.org/10.1093/brain/117.3.553

Figure Captions

Figure 1. Schematic of the task. The participant sees an instructional cue (left or right arrow) on the screen, which directs their attention to their left or right side. They then hear two phrases spoken by different talkers, which each contain a colour word and a number word: one phrase is presented on their left side and the other is presented on their right side. They are instructed to report the colour and number words from the side indicated by the instructional cue. In the example shown in the figure, the cue directs attention to the left side, and the correct answer is "White 1".

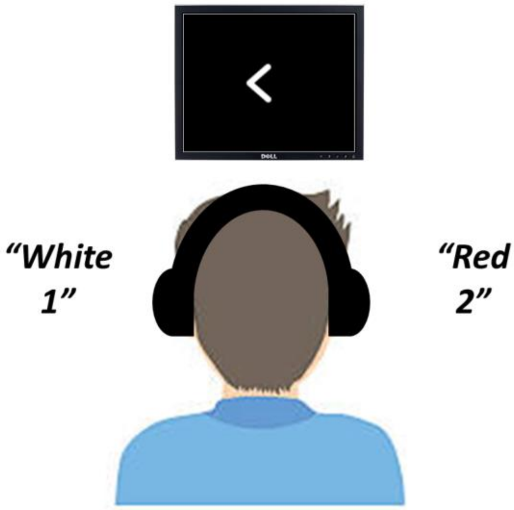


Figure 2. Schematic detailing the structure of a generative model of selective attention during cocktail party listening. States are displayed (in blue) on the upper row, and outcomes are displayed (in orange) on the lower row. Note that only 5 of the 17 Response states are illustrated in the figure, to avoid visual clutter. Grey solid arrows represent likelihood mappings between states and outcomes that have high probabilities, and grey dots at the intersection between two lines indicate that a state factor modulates the likelihood of an outcome under another state. For clarity, only one example set of mappings is shown in the figure—corresponding to a case in which *Spatial Attention* is Right, the *Target Colour Word* is "Red", and the *Target Number Word* is "1". Grey dashed arrows show mappings between word states and word outcomes that would have high probabilities if the *Spatial Attention* state was Left. The mapping between *Response* and *Feedback* is modulated by the *Target Colour* and *Target Number* factors (modulation not displayed). Transitions in the response factor depend upon the policy selected.

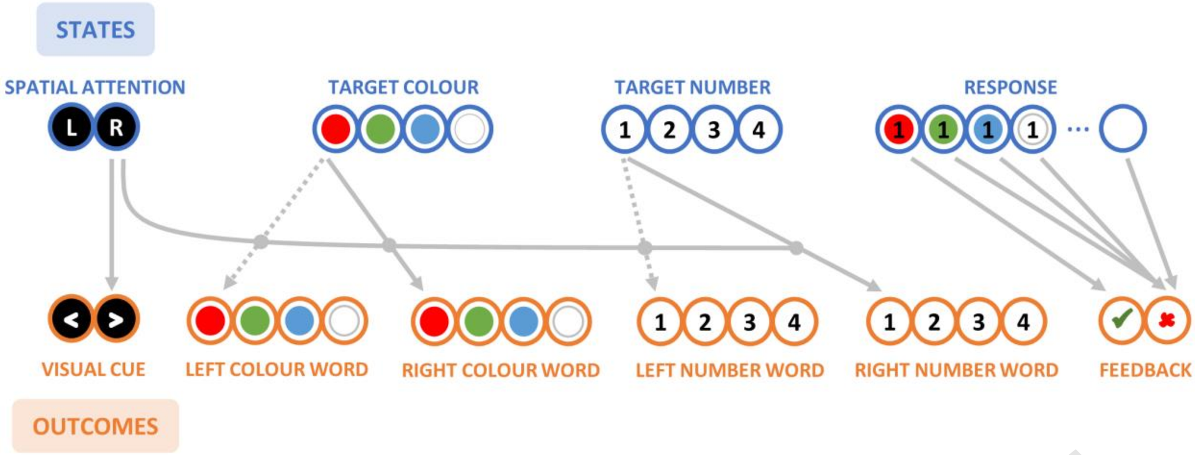


Figure 3. Details of the 4 different mechanistic hypotheses that produce characteristic errors. To test each hypothesis, a different likelihood precision was manipulated parametrically at 26 levels. This figure displays the range of precision parameter values, the default value of the precision parameter when other parameters were manipulated, and a schematic that illustrates the location of the manipulated precision in the model—which is indicated by the black arrows.

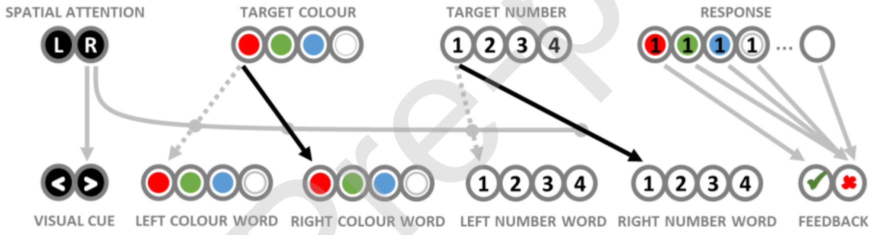
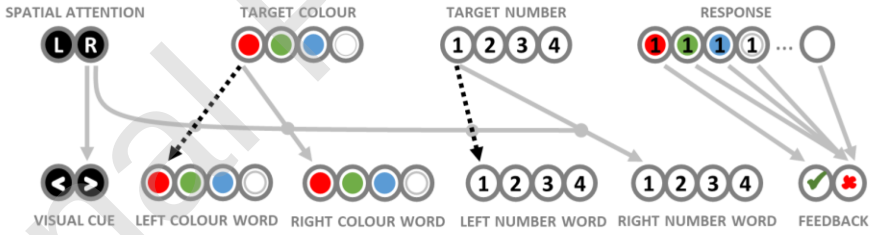
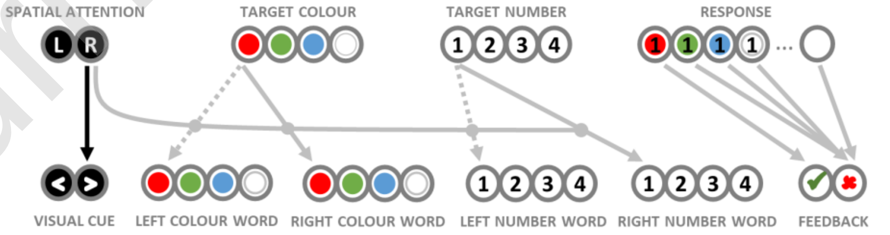
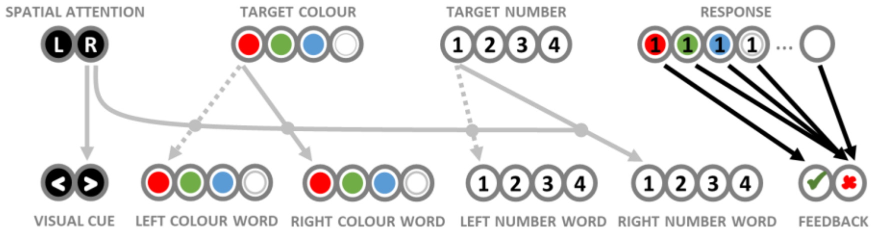
Precision parameter	Range	Default value	Schematic showing place in model
1	0–5	High (1024)	 SPATIAL ATTENTION L R TARGET COLOUR TARGET NUMBER 1 2 3 4 RESPONSE 1 1 1 1 ... VISUAL CUE LEFT COLOUR WORD RIGHT COLOUR WORD LEFT NUMBER WORD RIGHT NUMBER WORD FEEDBACK
2	0–1024	Low (0)	 SPATIAL ATTENTION L R TARGET COLOUR TARGET NUMBER 1 2 3 4 RESPONSE 1 1 1 1 ... VISUAL CUE LEFT COLOUR WORD RIGHT COLOUR WORD LEFT NUMBER WORD RIGHT NUMBER WORD FEEDBACK
3	0–5	High (1024)	 SPATIAL ATTENTION L R TARGET COLOUR TARGET NUMBER 1 2 3 4 RESPONSE 1 1 1 1 ... VISUAL CUE LEFT COLOUR WORD RIGHT COLOUR WORD LEFT NUMBER WORD RIGHT NUMBER WORD FEEDBACK
4	0–5	High (1024)	 SPATIAL ATTENTION L R TARGET COLOUR TARGET NUMBER 1 2 3 4 RESPONSE 1 1 1 1 ... VISUAL CUE LEFT COLOUR WORD RIGHT COLOUR WORD LEFT NUMBER WORD RIGHT NUMBER WORD FEEDBACK

Figure 4. Results from simulations. The left column shows the percent correct for 48 simulations, and the right column shows the error types. Errors are separated into three types: Masker, Mix, and Absent. Each error type is expressed as a percentage of all errors (i.e., based on the number of incorrect trials). (A) Effect of the precision of the likelihood mapping between beliefs about attended words and the observed words. (B) Effect of the precision of the unattended talker. (C) Effect of the precision of the likelihood mapping between *Spatial Attention* states and the observed *Visual Cue*. (D) Effect of the precision of the likelihood mapping between *Response* states and the *Feedback* outcome.

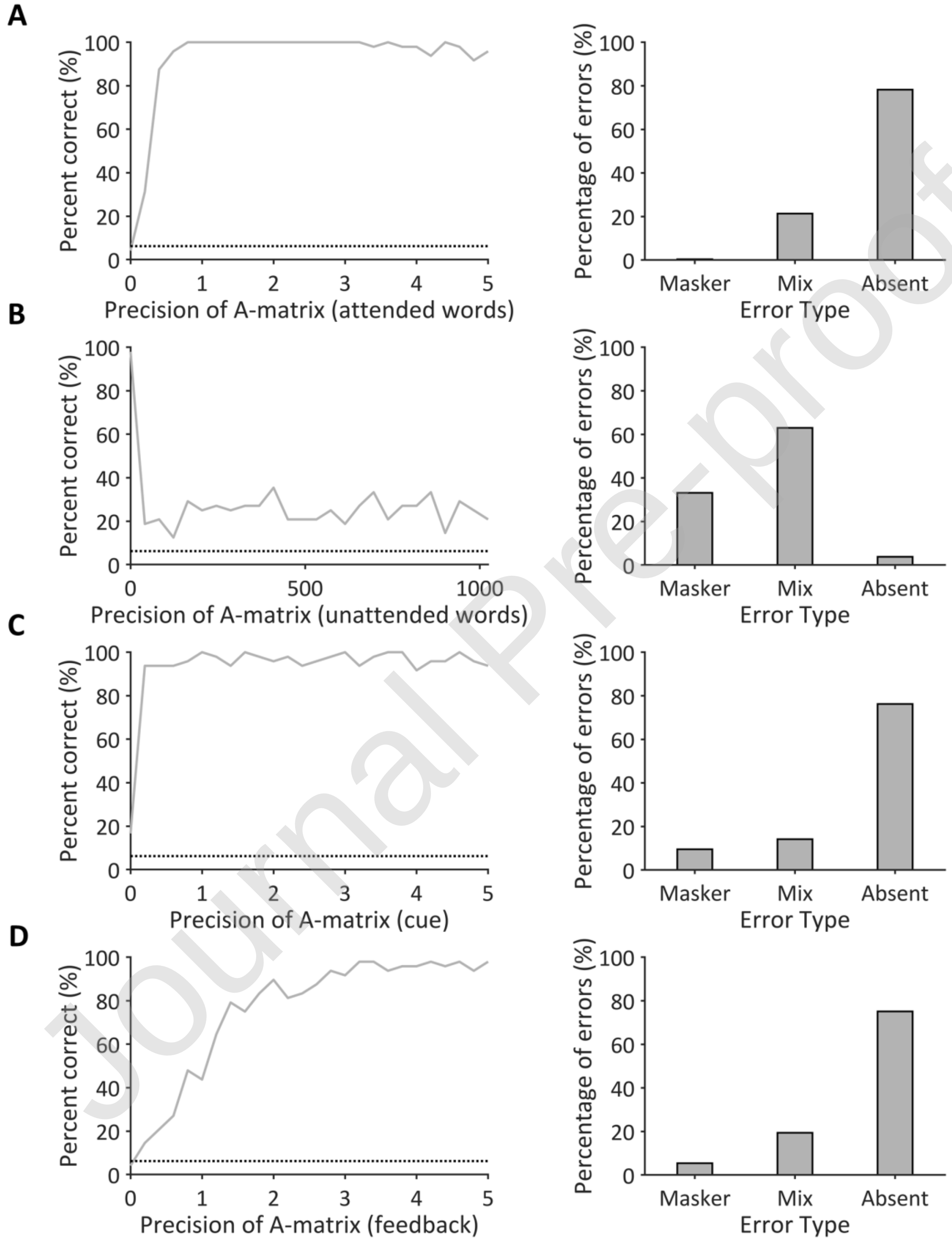


Figure 5. Combinations of A-matrix precision values for the mappings from beliefs about words (*Target Colour* and *Target Number* factors) to word outcomes on the attended and unattended sides

(*Left Colour Word*, *Right Colour Word*, *Left Number Word*, and *Right Number Word*). (A) Percent correct. (B) Masker errors as a percentage of all errors. (C) Mix errors as a percentage of all errors. (B) Absent errors as a percentage of all errors.

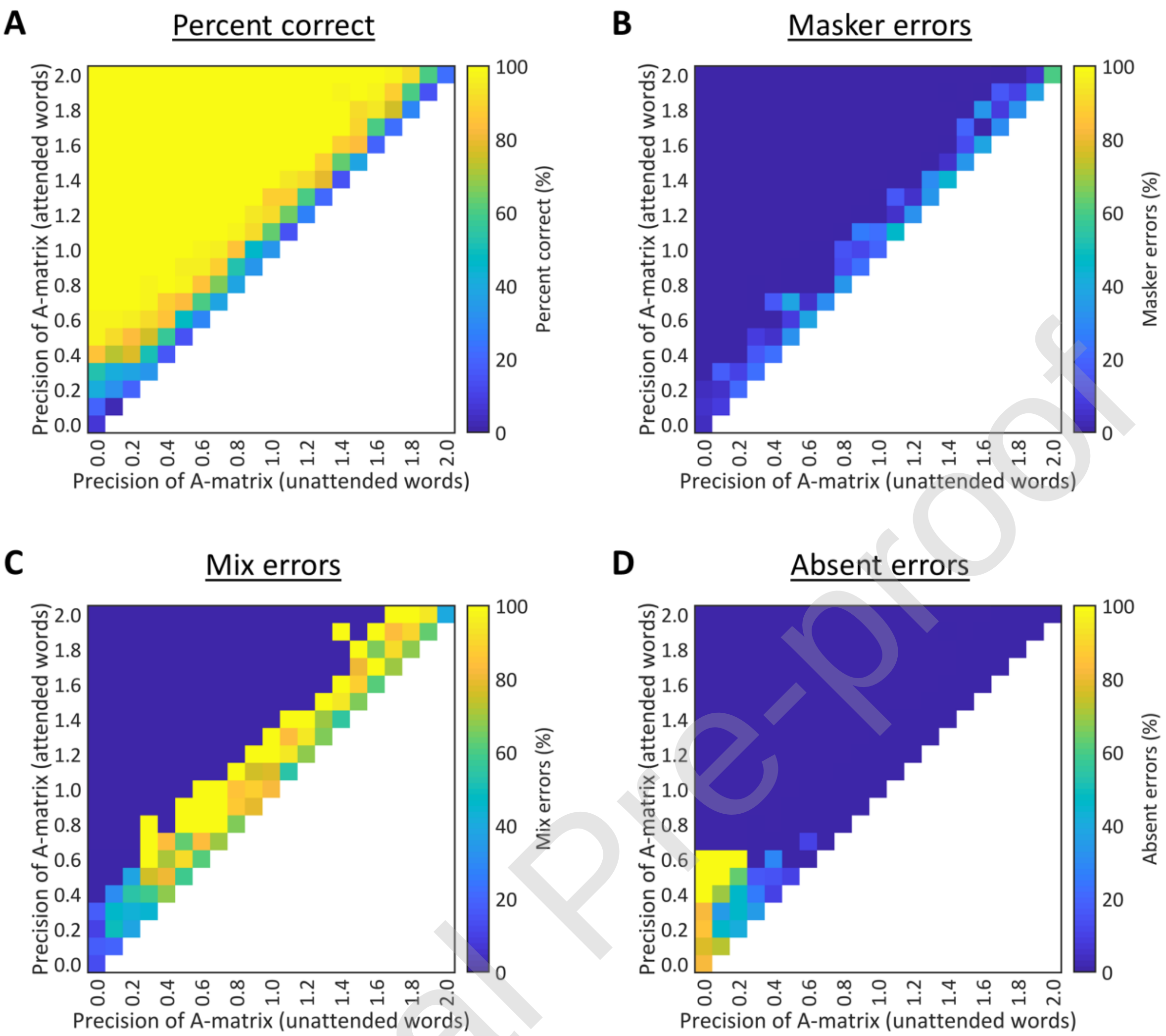
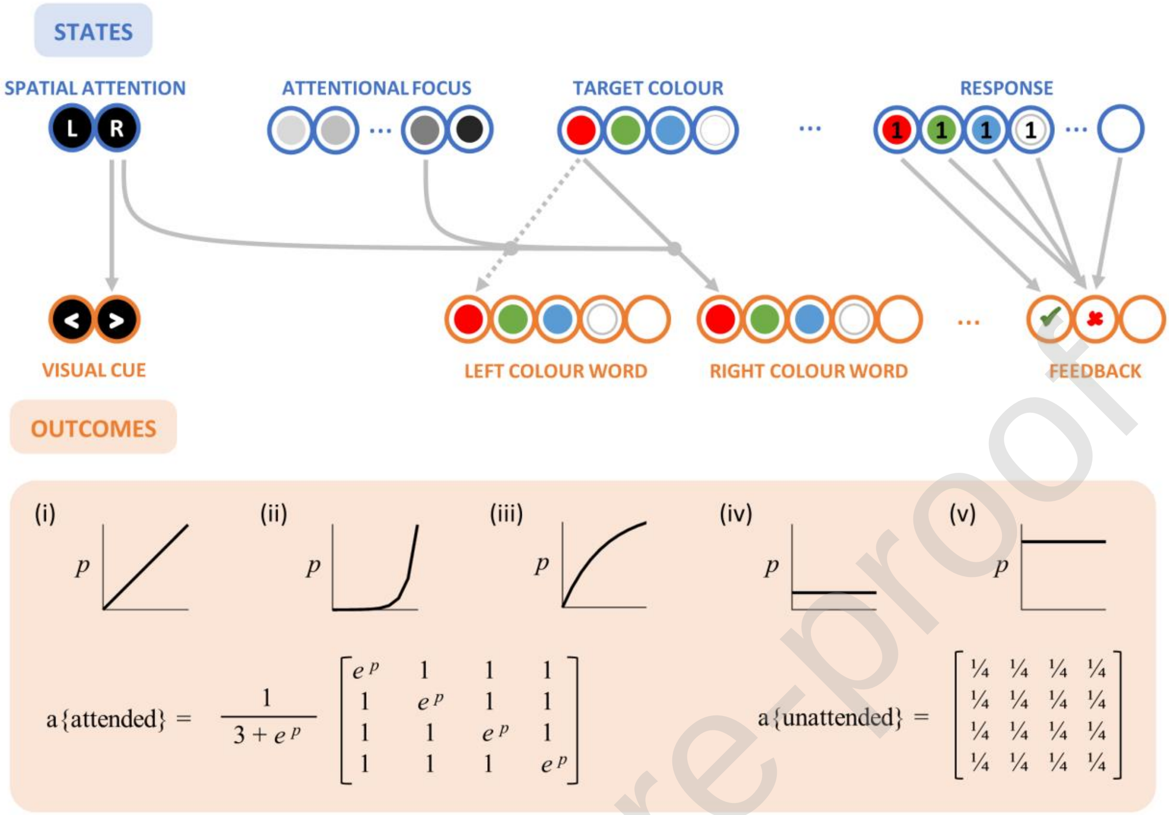


Figure 6. Schematic showing the generative model from Figure 2, with an additional *Attentional Focus* state factor that underwrites preparatory attention. States are displayed (in blue) on the upper row, and outcomes are displayed (in orange) on the lower row. For clarity, the *Target Number* states, *Left Number Word* outcomes, and *Right Number Word* outcomes are not displayed in the schematic, but act in the same way as the corresponding colour states and outcomes. Similarly, only 4 of the 10 *Attentional Focus* states and 5 of the 17 *Response* states are shown. Arrows represent likelihood mappings between states and outcomes that have high probabilities, and dots indicate that a factor modulates the likelihood mapping between a state and an outcome. The figure illustrates an example set of mappings for a trial in which the right talker was attended, who spoke the words “Red 1”. The lower (orange) panel shows how the *Attentional Focus* state affects the relevant mappings between states and outcomes. A precision parameter (*p*) is set by one of 5 distributions: panels (i)–(v) show the distributions from which this parameter is sampled under the (i) linear, (ii) exponential, (iii) exponential cumulative density, (iv) uniform with lower precision, and (v) and uniform with higher precision hypotheses. For attended outcomes (determined by the *Spatial Attention* state), the precision parameter becomes the temperature parameter in a softmax function, which determines the relationship between on-diagonal and off-diagonal elements. The likelihood mappings between the target word states and the *unattended* outcomes are uniform matrices (i.e., with zero precision)

under all five hypotheses. In terms of neuronal implementation, these likelihood mappings—and their time-dependent changes—correspond to effective connectivity, where short-term changes in underlying synaptic efficacy may be mediated by neuromodulatory synaptic mechanisms.



$$a\{\text{attended}\} = \frac{1}{3+e^{p}}\begin{bmatrix} e^{p} & 1 & 1 & 1 \\ 1 & e^{p} & 1 & 1 \\ 1 & 1 & e^{p} & 1 \\ 1 & 1 & 1 & e^{p} \end{bmatrix} \qquad a\{\text{unattended}\} = \begin{bmatrix} ¼ & ¼ & ¼ & ¼ \\ ¼ & ¼ & ¼ & ¼ \\ ¼ & ¼ & ¼ & ¼ \\ ¼ & ¼ & ¼ & ¼ \end{bmatrix}$$

Figure 7. Results from preparatory attention simulations, under different hypotheses about the shape of the temporal precision function. (A) Average reaction times for each combination of talker onset epoch and temporal hypothesis. The grey dotted line indicates the reaction time (RT) data from 20 human listeners, as reported in Holmes *et al.* (2018); for the purposes of this comparison, preparatory intervals of 0, 250, 500, 1000, and 2000 ms are assumed to correspond to talker onset epochs of 2, 3, 4, 6, and 10, respectively. Error bars indicate ± 1 standard deviation from the mean. The RTs simulated from the 5 models have been scaled (linearly) to the data reported in Holmes *et al.* (2018): this is why the simulated reaction times match the data at talker onset epochs of 2 and 10. (B) Root mean squared (RMS) errors between the (scaled) simulated RTs under each model and the RTs reported in Holmes *et al.* (2018). Model A: Linear; Model B: Exponential; Model C: Exponential cumulative density function (CDF); Model D: Uniform with lower precision; Model E: Uniform with higher precision.

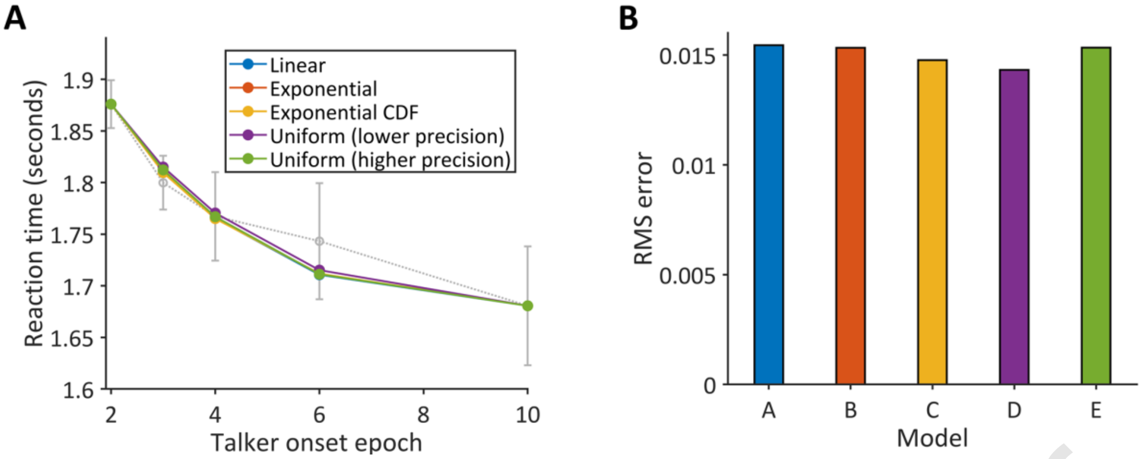


Figure 8. Examples of policy pruning under different talker onset values for the low-precision uniform model. The pattern of pruning was similar under all models. (A) The number of policies remaining at each epoch during the trial, starting when the visual cue was presented. (B) The number of policies remaining in each epoch, relative to the talker onset epoch. Within each plot, each line reflects a different Talker Onset Epoch condition.

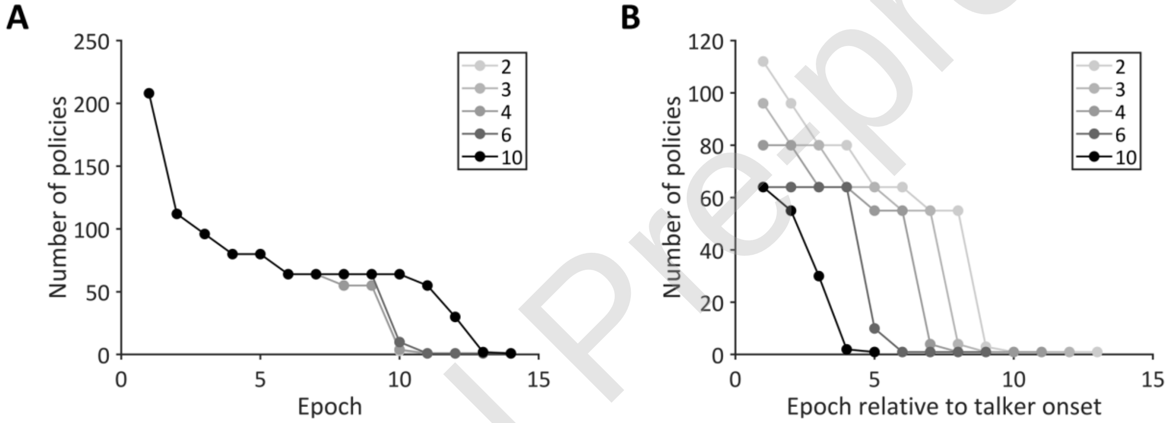
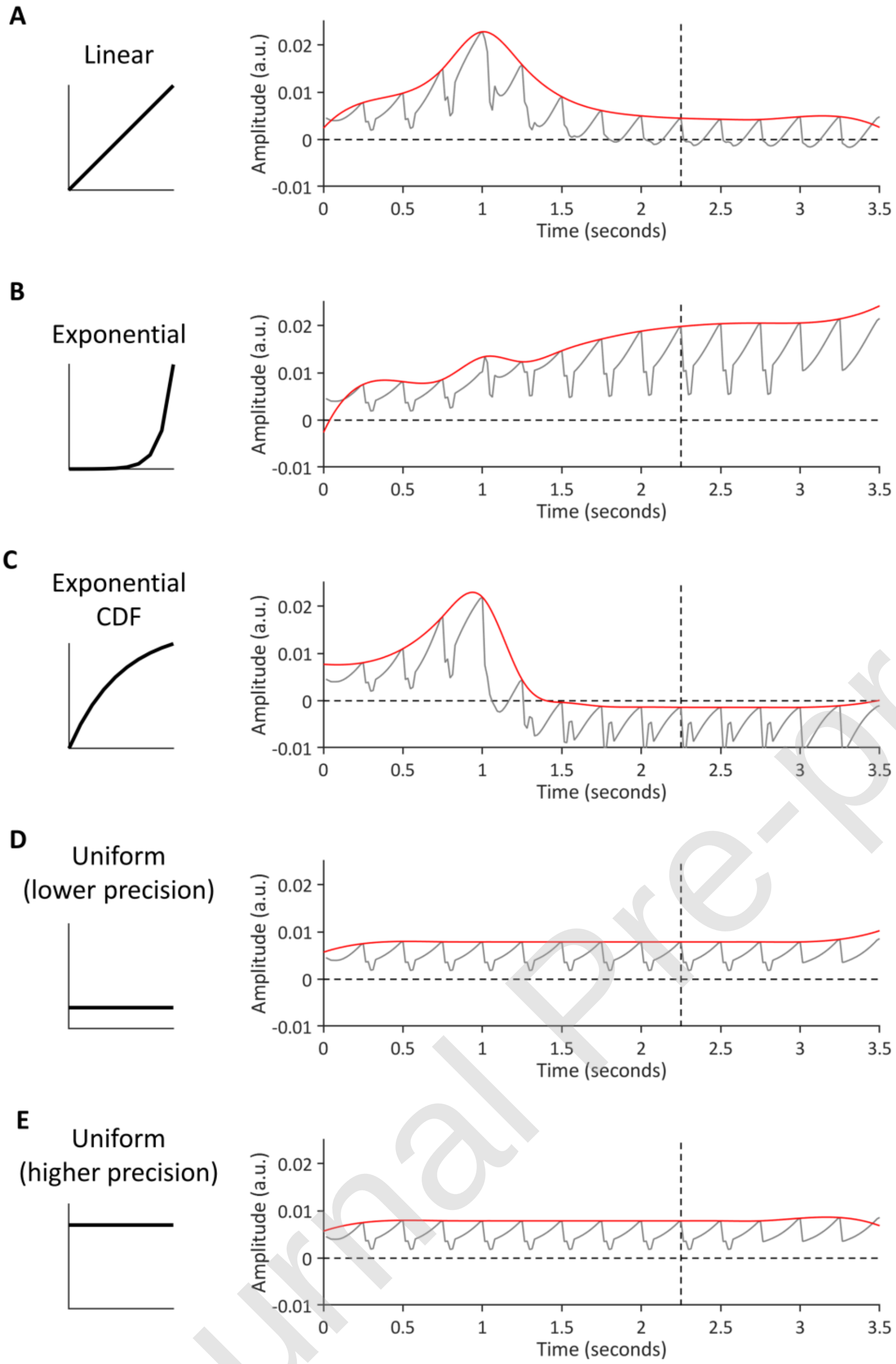


Figure 9. Examples of synthetic electrophysiological responses under different models. Grey lines show simulated electrophysiological responses, and red lines show the slower temporal pattern of simulated responses (for visualisation only). Each epoch is plotted with 0.25 seconds duration. On these plots, the presentation of the visual cue occurs at 0 seconds, and the talker onset corresponds to 2.25 seconds (as indicated by the dashed vertical lines). (A) Linear function. (B) Exponential function. (C) Exponential cumulative density function (CDF). (D) Uniform function with lower precision. (E) Uniform function with higher precision.


A
Linear
B
Exponential
C
Exponential
CDF
D
Uniform
(lower precision)
E
Uniform
(higher precision)
Amplitude (a.u.)
0.02
0.01
0
-0.01
0
0.5
1
1.5
2
2.5
3
3.5
Time (seconds)