scientific reports

Check for updates

OPEN

Simulating lesion-dependent functional recovery mechanisms

Noor Sajid[1]✉, Emma Holmes[1], Thomas M. Hope[1], Zafeirios Fountas[1,2], Cathy J. Price[1] & Karl J. Friston[1]

Functional recovery after brain damage varies widely and depends on many factors, including lesion site and extent. When a neuronal system is damaged, recovery may occur by engaging residual (e.g., perilesional) components. When damage is extensive, recovery depends on the availability of other intact neural structures that can reproduce the same functional output (i.e., degeneracy). A system's response to damage may occur rapidly, require learning or both. Here, we simulate functional recovery from four different types of lesions, using a generative model of word repetition that comprised a default premorbid system and a less used alternative system. The synthetic lesions (i) completely disengaged the premorbid system, leaving the alternative system intact, (ii) partially damaged both premorbid and alternative systems, and (iii) limited the experience-dependent plasticity of both. The results, across 1000 trials, demonstrate that (i) a complete disconnection of the premorbid system naturally invoked the engagement of the other, (ii) incomplete damage to both systems had a much more devastating long-term effect on model performance and (iii) the effect of reducing learning capacity within each system. These findings contribute to formal frameworks for interpreting the effect of different types of lesions.

Most patients who suffer functional impairments after brain damage improve over time. This has been demonstrated in motor[1–3], visual[4,5] and language[6,7]. The degree of functional recovery is highly variable and lesion-dependent[8–10]. Here, we distinguish between two distinct lesion-dependent recovery mechanisms. For a premorbid system with partial damage, recovery may entail re-learning that increases the functional capacity of the damaged system—and may involve peri-lesional activity[11]. In a severely damaged system, recovery depends on whether an alternative system can be used to reproduce the same functional output[12,13].

The ability to use alternative systems for the same task is referred to as degeneracy[14,15]. For example, when reading written words aloud, with regular spelling to sound relationships, sounds can be generated either by using learnt spelling-to-sound relationships or by whole word recognition. If the predominant word recognition system fails, readers can instead use spelling-to-sound relationships. At a neurological level, degeneracy supports functional recovery, following damage to components of the premorbid system, by enabling intact cortical regions and pathways to recapitulate the same function. The new system may engage distinct components, relative to the premorbid system, but may share undamaged components[13]. Like the reading example, different components may not have the same function in isolation, but together can produce the same output as the premorbid system. Degenerate systems may therefore support better behavioural performance than relying on a partially damaged premorbid system. However, when the lesion is extensive or affects multiple areas; all available degenerate systems may also be rendered dysfunctional.

If two neural systems for the same function are equally efficient and sufficient, then switching from one system to another could occur instantaneously. Conversely, if one system is preferred over another then it takes time to attain premorbid levels of proficiency—and may require functional reorganisation mediated by experience-dependent plasticity: i.e., re-learning due to behavioural experience[16–20]. During re-learning, experience-dependent plasticity helps to restore and compensate for functional deficit[21]. These learning processes depend on multiple mechanisms manifest at different temporal scales: namely, short-term plasticity mechanisms that occur rapidly (e.g., neuromodulatory changes in synaptic efficacy) and slower long-termreorganisation that support functional recovery (e.g., long-term potentiation and depression)[20]. These might be complemented by changes in the neural structure[22] and/or electrophysiology[23]. Previous models of plasticity-related recovery have shown re-learning can retune and realise the contribution of peri-lesional regions[12,24].

Using in-silico lesions in a computational model of word repetition (i.e., repeating heard words), we characterise two functional recovery mechanisms. The first is the rapid recruitment of an alternative system that can

[1]Wellcome Centre for Human Neuroimaging, University College London, UCL Queen Square Institute of Neurology, 12 Queen Square, London WC1N 3AR, UK. [2]Huawei 2012 Laboratories, London, UK. ✉email: noor.sajid.18@ucl.ac.uk

reproduce the same outcome (in the context of degeneracy); the second is long-term plasticity within either the premorbid or alternative system. The recovery mechanism triggered is expected to depend on the structural, and implicitly computational, resources available to perform the task. We modelled these processes using active inference, which treats perception and action as belief updating under a particular generative model of the environment. Central to this approach is the notion that behaviour is Bayes optimal under some prior beliefs (following the complete class theorem). This allows us to characterise patients with brain damage as operating under (possibly lesional) priors that constitute their generative model; priors are statistical contingencies encoded by model parameters (e.g., synaptic connection strengths). Additionally, active inference provides a formal way of measuring: synaptic connectivity changes during learning, the accompanying neurophysiology[25] and degeneracy. Here, degeneracy is the entropy of Bayesian beliefs about the causes of sensations and redundancy is a measure of how much beliefs need to change to explain current observations[26].

Building on prior work[26,27], we used a simulated work repetition paradigm to measure degeneracy, redundancy, and task performance under four different levels of in-silico lesion severity. The key extension in our current model of word repetition, compared to our previous model, is that we introduced two distinct degenerate networks—premorbid and alternative—into a hierarchical model that also captured the neuromodulatory aspect of attention. This allowed us to simulate the effects of different types of lesions that were not investigated in our prior work. The premorbid network represents the preferred (i.e., predominant) circuitry, prior to lesioning[28,29]. The alternative network is less experienced but has the capacity to learn.

The resulting hierarchical model was lesioned in four ways: first, we simulated changes in model performance when the premorbid system was completely disconnected, and output depended on the less experienced alternative system. Second, we repeated Lesion 1 but added a second lesion to the intrinsic connectivity of the alternative system. This limited recovery by reducing the capacity of the alternative system to learn, over time, via experience-dependent plasticity. Third, rather than completely disconnecting the premorbid system, as in Lesions 1 and 2, we partially disconnected both the premorbid and alternative systems. This made it difficult for the higher level to identify which context (premorbid or alternative) would be most effective. Finally, we repeated Lesion 3 (partial disconnections to both systems) with an additional lesion to reduce relearning capacity.

Below, we briefly introduce active inference, generative models and the types of in-silico lesions that can be induced. In the following sections, we describe the word repetition model, how we use it to simulate functional recovery mechanisms and the results. Finally, to make quantitative and disambiguating predictions for future empirical work, we simulated electrophysiological responses that have already been associated with a decrease in baseline neuronal firing[23,30], and attenuated post-synaptic sensitivity[31,32].

Active inference

Active inference postulates that the brain self-organises by optimising two complementary objectives: (1) fitting the model to (sampled) observations to minimise variational free energy (F; surprisal)[25,33,34] and, (2) selecting actions that minimise expected free energy (G; uncertainty)[35,36]. The variational free energy is the complexity cost incurred in forming accurate posterior beliefs about causes of sensation:

$$F = \underbrace{D_{KL}[Q(s)||P(s)]}_{\text{complexity}} - \underbrace{\mathbb{E}_{Q(s)}[\log P(o|s)]}_{\text{accuracy}} \tag{1}$$

and, expected free energy:

$$G \approx -\underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(Q(s_\tau|o_\tau, s_{\tau-1}, \pi)) - \log(Q(s_\tau|\pi)\right]}_{\text{mutual information}} - \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(P(o_\tau))\right]}_{\text{expected log evidence}} \tag{2}$$

where $\tilde{Q} = P(o_\tau|s_\tau)Q(s_\tau|\pi)$ and $Q(o_\tau|s_\tau, \pi) = P(o_\tau|s_\tau)$.

Using these objectives, expectations about hidden states, policies and precision are optimized through inference, and model parameters optimized through learning. This involves the variational message passing of sufficient statistics of posterior beliefs (i.e., expected probability) among neuronal populations. Note, the variational message passing can be formulated as a gradient descent on variational free energy[37,38].

Deep temporal models. The process theory underwriting active inference is based on a partially observable Markov decision process (POMDP). This can be defined as a generative model with discrete outcomes that are caused by discrete hidden states[37,39]. These models can have a deep temporal (i.e., hierarchical) structure, where the outcomes of one level generate the hidden states at the level below[37,39]. The equation below defines the factorized form of a deep temporal model, via its joint probability over outcomes and hidden states.

$$P(\tilde{o}, \tilde{s}, \pi) = \prod_{i=1}^{L} \prod_{\tau'} \prod_{\tau} \underbrace{P\left(o_\tau^{(i)}|s_\tau^{(i)}\right)}_{Cat\left(\mathbf{A}^{(i)}\right)} \underbrace{P\left(s_1^{(i)}|s_{\tau'}^{(i+1)}\right)}_{Cat\left(\mathbf{D}^{(i)}\right)} \underbrace{P\left(s_{\tau+1}^{(i)}|s_\tau^{(i)}, \pi^{(i)}\right)}_{Cat\left(\mathbf{B}_{\pi,\tau}^{(i)}\right)} \sigma\left(-\mathbf{G}\left(\underbrace{P\left(o_\tau^{(i)}|s_{\tau'}^{(i+1)}\right)}_{Cat\left(\mathbf{C}_\tau^{(i)}\right)}\right)^{(i)}\right) \tag{3}$$

where $\tilde{o}$ and $\tilde{s}$ denote sequences of outcomes and hidden states respectively, until the current time point: i denotes the i-th hierarchical level and L the total number of levels and policies,π are trajectories over potential action space.

Here, outcomes depend upon hidden states and hidden states depend upon policies. Outcomes and hidden states are parametrized by two distinct categorical distributions:$A^{(i)}$ and $B^{(i)}_{\pi,\tau}$. $A^{(i)}$ is the likelihood function, that maps hidden states to outcomes at the i-th level, and $B^{(i)}_{\pi,\tau}$ the distribution that maps transitions from one hidden state to the next, at the i-th level, under policy π. Successive levels of the generative model are linked by $D^{(i)}$, which defines the mapping between hidden states at $i+1$-th level to the initial states at the i-th level below.

Policies, and the next action, are selected by sampling from the softmax function of expected free energy. Consequently, policies are more probable, a priori, if they minimise expected free energy. Using this, hidden state sequences are generated using $B^{(i)}_{\pi,\tau}$ determined by the selected policy. These hidden states generate outcomes and initial hidden states in the level below (according to $A^{(i)}$ and $D^{(i)}$). They influence the expected free energy through $C^{(i)}$ and the policies that determine transitions among subordinate states. Here, the model parameters, i.e.,$A^{(i)}$, $D^{(i)}$, $C^{(i)}$&$B^{(i)}_{\pi,\tau}$ are equipped with a prior (categorical) distribution. The key aspect of this generative model is that transitions among states proceed at different rates at different levels of the hierarchy. Essentially, hidden states at higher levels contextualize trajectories of hidden states at lower levels, enabling a deep dynamic narrative[37].

This generative model can be regarded as the 'structure' in structure–function relationships and is thought to underwrite functional brain architectures that realize active inference. This perspective allows us to associate *in-silico* lesions with anatomical lesions. Specifically, when defining model inversion in terms of message passing, $A^{(i)}$ can be regarded as extrinsic connections i.e., between different regions/neuronal populations while $B^{(i)}$ can take the form of intrinsic connections i.e., within and between the cortical layers of a single region[40].

Learning. Each synthetic subject has implicit prior beliefs about their model parameters (i.e., likelihood -$A^{(i)}$, transitions – $B^{(i)}$, etc.). This includes prior beliefs over model parameter priors (i.e., hyperpriors), which are learned through Bayesian belief-updating[25,41]. The natural choice for the conjugate hyperprior over categorical priors is a Dirichlet distribution. This means that hyperpriors can be expressed simply in terms of Dirichlet concentration parameters[42], which represent each state-to-outcome (for $A^{(i)}$) and state-to-state ($B^{(i)}$) mappings or connectivity:

$$P(A^i|a^i) = Dir(a^i) \Rightarrow \begin{cases} \mathbb{E}_{P(A^i|a^i)}\left[A^i_{ij}\right] = \dfrac{a^i_{ij}}{\sum_k a^i_{kj}} \\ \mathbb{E}_{P(A^i|a^i)}\left[\log A^i_{ij}\right] = \psi(a^i_{ij}) - \psi\left(\sum_k a^i_{kj}\right) \end{cases}$$
$$P(B^i|b^i) = Dir(b^i) \Rightarrow \begin{cases} \mathbb{E}_{P(B^i|b^i)}\left[B^i_{ij}\right] = \dfrac{b^i_{ij}}{\sum_k b^i_{kj}} \\ \mathbb{E}_{P(B^i|b^i)}\left[\log B^i_{ij}\right] = \psi(b^i_{ij}) - \psi\left(\sum_k b^i_{kj}\right) \end{cases} \quad (4)$$

where ψ is the digamma function.

The Dirichlet parameters can be thought of as 'pseudo-counts' i.e., as observations are made, they accumulate Dirichlet parameters that best assimilate sensory data. The more often a given pairing (of state and outcome, or past and present state) is observed, the greater the number of counts attributed to that pairing[43,44]. Thus, beliefs change to a greater extent when the counts are low—because few pairings have been observed—than when they are high. This accumulation process closely resembles Hebbian plasticity[45], where synaptic efficacy is reinforced upon the simultaneous firing of a pre and postsynaptic neuron[46]. This learning process affords experience-dependent plasticity: namely, the strengthening of synaptic connections during belief updating.

Simulating in-silico lesions. In-silico lesions can be simulated by manipulating precision, where precision scores confidence (i.e., the inverse of uncertainty) in beliefs about the causes of sensations. As in our prior work[26,27], we manipulate precision, over model parameters $A^{(i)}$ and $B^{(i)}$ which results in different types of damage that can be linked to pathological lesions in the human brain.

Precision over $A^{(i)}$ (sensory precision) corresponds to the confidence with which the model can infer the cause of observations on the basis of prior experience. Decreasing the precision over $A^{(i)}$ mimics a lesion to extrinsic connections (i.e., between regions expectations about causes and observations). This results in uncertainty about the causes of observations. Extending our prior work[26], we distinguish between structural and functional disconnections. When precision over $A^{(i)}$ is completely imprecise (i.e., probability distribution is uniform) we induce a "structural disconnection" between the observations and their underlying causes, meaning that it is not possible to resolve the causes. A less severe lesion (i.e., slightly imprecise distribution), in contrast, only results in a "functional disconnection" leaving imprecise, ambiguous relationships between causes and outcomes, that can, in principle, be resolved through re-learning.

Precision over $B^{(i)}$(state transition precision) corresponds to confidence with which the model can predict the present from the past (i.e., infer state transitions using prior transition probabilities). Decreasing precision over $B^{(i)}$ induces a focal lesion to intrinsic connections (e.g., within and between the cortical layers of a single region). This precision affects the model's ability to learn (i.e., limiting the ability to update beliefs about causes). As with lesions to precision over $A^{(i)}$, lesions to precision over $B^{(i)}$ can either be structural (completely imprecise) or functional (slightly imprecise). Structural lesions prevent appropriate learning and belief updating. Functional lesions reduce the capacity to learn and update beliefs.

Belief updating is also compromised when sensory precision (i.e., over $A^{(i)}$) is reduced. This can be linked to neuromodulatory control when precision is implicated at the lower level in deep temporal models. In other words,

the higher level must appropriately disambiguate between similar competing lower-level causes and modulate their precision. For example, acetylcholine release, in extrinsic connectivity, is thought to increase sensory precision (e.g., by boosting bottom-up sensory signals) enabling the brain to respond optimally[47]. Lowering precision over $A^{(1)}$ is therefore expected to impair neuromodulatory control of synaptic efficacy in intrinsic connections.

Simulating electrophysiological measurements. The form of the (variational) message passing mandated by active inference, as shown below, allows us to associate variables with idealised electrophysiological measurements[25,40]:

$$\begin{aligned}
\log s_{\tau\pi^n} &= v_{\tau\pi^n} - \log z_{\tau\pi^n} \\
\dot{v}_{\tau\pi^n} &= \varepsilon_{\tau\pi^n} \\
z_{\tau\pi^n} &= \sum \exp\left(v_{\tau\pi^n}\right) \\
\varepsilon_{\tau\pi^n} &= \log A_\tau s_{\tau\pi\setminus n} \cdot o_\tau + \frac{1}{2}\left(\log\left(B_{(\tau-1)\pi^n} s_{(\tau-1)\pi^n}\right) + \log\left(B^{\dagger}_{(\tau+1)\pi^{n^2}} s_{(\tau+1)\pi^n}\right)\right) - \log s_{\tau\pi^n}
\end{aligned} \tag{5}$$

Here, $s^n_{\tau\pi}$ denotes the expected hidden state factor (n) at time (τ), conditioned on a policy (π); z is the partition function. v and ε are auxiliary variables that play the roles of membrane depolarization (i.e., post-synaptic potential) and prediction error, respectively. v is computed from the inputs of other neurons and transformed to s via the partition function,z. Note, that s is the signal that is propagated to other neuronal populations and is analogous to firing rate, as measured by single unit recordings. The rate of change of v can be associated with local field potentials, after bandpass filtering between 4 and 32 Hz[25].

A generative model of word repetition

To illustrate how particular lesions could trigger different recovery mechanisms, we extended our previous generative model[26,27] and active inference scheme for simulating word repetition[24,48,49]. The subject (i.e., model) hears a single spoken word and must repeat it. If repeated correctly, they receive positive feedback (and negative otherwise).

Previous model of word repetition. In the previous version of the model, there were three state factors (Target Word, Spoken Word, Epoch) and three outcome modalities (Proprioception, Word, and Evaluation). The Target Word factor has five states, corresponding to the words that could be heard: red, blue, table, triangle and square. The Spoken Word factor contains beliefs about what word should be repeated: red, blue, table, triangle and square. The Epoch factor codes two different stages of the trial: listening to a target word (epoch 1), repeating it and receiving evaluation (epoch 2). In terms of outcomes, the Proprioception outcome reports whether the subject moved their mouths to speak or not. This depends on the state of the epoch factor: if the subject believes they are in epoch 1 (listening to the target word), then they are not speaking but if they believe they are in epoch 2 (repeating the target word), then they are speaking. The Word outcome reports the heard word (target or spoken): red, blue, table, triangle or square. This depends on the states of the Target Word, Spoken Word and Epoch factors. If the subject believes they are in epoch 1, the word outcomes depend on the Target Word factor but if the subject believes they are in epoch 2, the word outcomes depend on the Spoken Word factor. Finally, the Evaluation outcome indicates if the spoken word was the same as the target word. The outcome depends on the Epoch factor: positive (correct) if the spoken word is the same as the target word at epoch 2, negative (incorrect) if the spoken word is not the same as the target word at epoch 2, or neutral (during epoch 1). For a discussion of the functional architecture entailed by this generative model—and the underlying neurobiology—please see Sajid et al.[26].

Current (hierarchical) model of word repetition. The current model extended our previous model (described above) in two ways. First, the new model is hierarchical with two levels (Fig. 1). The top level (not included previously) has one state factor: network. The Network factor contains two states: premorbid or alternative. The premorbid system is associated with greater precision than the alternative system—as if the model were in an attentive compared to an inattentive state, respectively. This follows because changes in precision are generally associated with attention[50,51].

The second extension was the addition of a fourth state factor (Context) to the lower level that serves as the target of top-down messages from the hierarchical level above. The Context factor denotes attentive (precise) or inattentive (imprecise) beliefs depending on the Network factor at the higher level: if the subject believes they are using the premorbid system, then they will be in the attentive context, and otherwise inattentive. A shift in context from attentive to inattentive at the lower level reduces the precision of the mappings between other state factors and outcomes i.e., decreased confidence about the causes of sensations. Notice that this deep model equips our synthetic subject with a more nuanced and context sensitive processing capacity that can be likened to having an attentional set, which is sensitive to—and contextualises—processing at lower levels.

The context factor is particularly important for the current simulations because it effectively duplicates the message passing under the remaining factors (c.f., premorbid and alternative systems). These context states are distinguished only in terms of the sensory precision in the mappings between states and outcomes. Therefore, when precision in these mappings is similar for both contexts, the higher (Network) level cannot infer the appropriate attentional set or context. The Network factor at the higher level is only connected to the Context factor at the lower level. This setup enables beliefs about the lower-level Context factor to be updated if there is a change

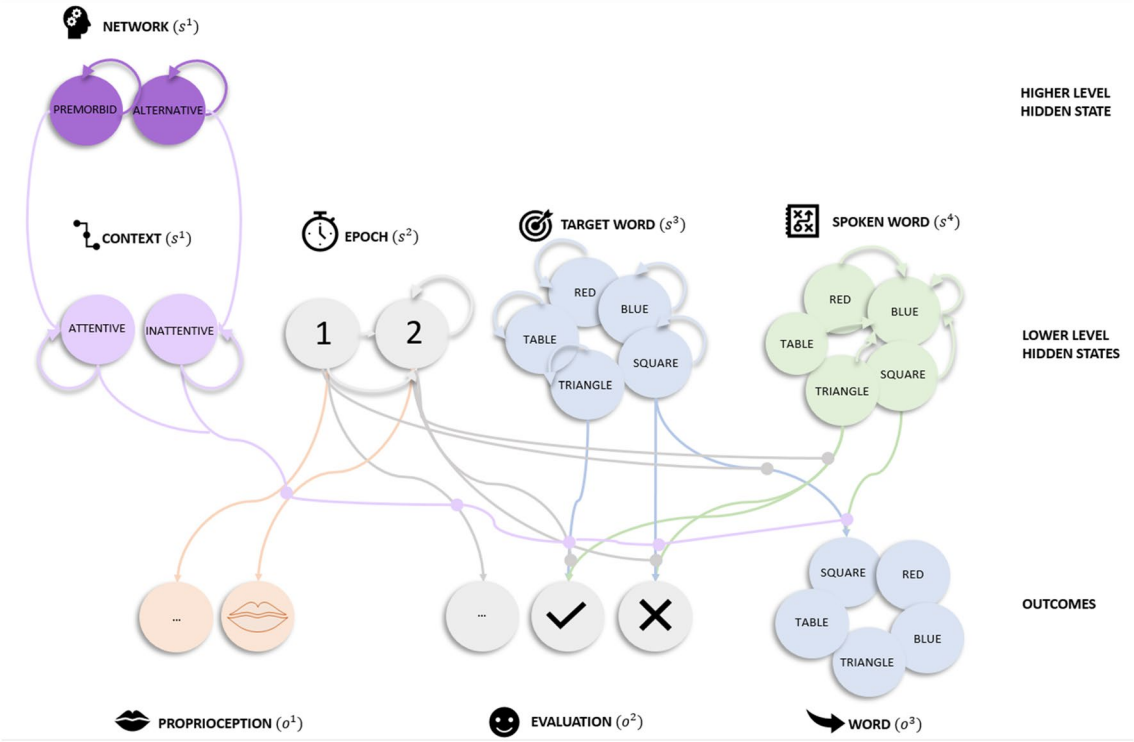

Figure 1. Generative Model. Graphical representation of the generative model for word repetition. The model comprises two levels. There is one higher-level state factor (Network), four lower-level state factors (Epoch, Target Word, Spoken Word, and Context) and three outcome modalities (Proprioception, Evaluation and Word). *Network* (2 states) at the higher level specifies the system in play (premorbid or alternative). *Epoch* (2 states) indexes the phase of the trial. During the first epoch, the target word is heard. The second epoch involves repeating the word and concomitant evaluation. A positive evaluation is provided if the Word outcome matches the Target Word state, and a negative evaluation otherwise. The *Target Word* factor (5 states) lists the words the experimenter can ask the participant to repeat. The *Spoken Word* factor includes the words that the model can choose to say (5 states). The *Context* factor (2 states) specifies whether the subject is in an attentive or inattentive state. Lines from states to outcomes represent the likelihood mappings and lines mapping states within a factor represent allowable state transitions. The lines represent plausible connections, and their absence reflects implausible connections. To avoid visual clutter, only a few likelihoods and transition probabilities are shown, but they are consistent across the different state factors and outcome modalities. For example, the Word likelihood mapping from the Target Word state (square) to the Word outcome (square) is shown for Epoch 1, but similar mappings apply when mapping between the blue Target Word and blue Word outcome and between the triangle Target Word state and triangle Word outcome, etc. One (from a total of 5) example transition probability is highlighted for the Spoken Word state, i.e., the transition is always to blue, regardless of previously spoken word (red, table, triangle, square or blue). This transition represents the choice to say 'blue'. Similar mappings are applied when choosing to say 'triangle'. Actions then correspond to the selection of particular transition probabilities. The Context factor modulates the strengths of the likelihood mappings between the other factors and outcomes, depending on whether the model is in an attentive or inattentive state.

in beliefs about the Network at the higher level. All of the other state factors at the lower level are conditionally independent of the higher level—and are exactly the same as on our previous model.

Model parameters. Figure 1 illustrates how States are connected to Outcomes using lines or edges. The strengths of these connections are defined by the likelihood parameters, $A^{(i)}$ which indicate the sensory precision of extrinsic connections ("Simulating in-silico lesions" above). Each state also has a transition matrix, $B^{(i)}$ (denoted by arrows in Fig. 1) that maps the state at the current timepoint to the next time point. This indicates how precisely the current state will transition to another state (via intrinsic connectivity).

For the Spoken Word factor, there are five transitions. These transitions depend upon the action that the subject selects: upon taking an action, this state always transitions to the selected word. For example, when the subject chooses to say 'blue', regardless of previous the word (i.e., red, triangle, etc.), the Spoken Word state will be blue (highlighted in Fig. 1). For the lower-level Epoch factor, states transition from epoch 1 to epoch 2 which is an absorbing state (final epoch). For the Context factor and the Target Word factor, transitions were represented in an identity matrix (of size two and five). This means these factors stay the same throughout a trial. The transitions were also represented with an identity matrix for the higher-level network factor.

Lesion	Sensory precision A^1 in premorbid system	Sensory precision A^1 in alternative system	Target word transition B^1 precision	Dirichlet count
0	1.0	0.9	1.0	1000
1	0.0	0.9	1.0	1000
2	0.0	0.9	0.4	1
3	0.1	0.1	1.0	1000
4	0.1	0.1	0.4	1

Table 1. Overview of precision manipulation. The table presents a breakdown of the five models and the underlying changes in precision that simulated particular lesion types—and the initial Dirichlet counts. 1 denotes high precision, 0 low precision (i.e., a uniform distribution) and everything else a gradation between the two.

The model had strong preferences for receiving positive evaluation (i.e., getting the repetition correct) and strong aversion to negative evaluations. It was also equipped with 5 different one-step-ahead polices (action trajectory) to choose from, corresponding to each word that could be spoken. We specified the following hyper-parameters: threshold for passing back control to the higher-level was $e^{-16} \approx 0$ and the decaying learning rate was 2.

Relationship to other models. The word repetition model was implemented as a partially observable Markov decision process and some parallels can be drawn between this and existing models. Our model has two parallel contexts, defined by the higher-level state factor, which specifies a premorbid and alternative system. This equips the model with two-way (albeit asymmetric) processing, due to imbalanced (attentional) resource allocation, that may be involved in word repetition. The asymmetry here is that the subject gives prior preference to attentive (i.e., precise), over inattentive (i.e., imprecise) beliefs. In contrast, a symmetrical model would select attentive or inattentive states, with equal probability. This makes our model formally similar to the recurrent neural network presented in[52], with dual structures that support function, but are asymmetrical due to differences in computational resource allocation.

Our Bayesian approach is formally distinct from some previous (cognitive) models of word repetition; for example, recurrent[24,52] and adaptive[53,54] neural networks. These formulations can be regarded as deterministic function approximators, after an initial training phase. In contrast, our approach does not require 'training' and can be used to simulate outcomes based on a (learnt or prespecified) probability distribution or belief state.

Our model has close ties to the state feedback control model of speech production using Kalman filtering, another Bayesian algorithm[55]. The Kalman filter is the Bayes-optimal solution under a linear generative model, but a cascade of such solutions would not be the optimal solution to (non-linear and non-Markovian) hierarchical models[56]. Conversely, active inference considers the hierarchical system as a whole and provides an optimal solution in the form of variational (Bayesian) filtering or belief updating.

Simulating the effect of in-silico lesions to our word repetition model

Control simulation (Lesion 0). To measure differences in recovery, we simulated a control model without any lesion. This acts as a sanity check for the architecture of the premorbid system and simulates the behaviour of healthy subjects performing the task. At the top level, the premorbid system was given greater precision than the alternative system (see Lesion 0, Table 1). This increased attention to the premorbid system at the lower level[50,51] resulting in strong prior expectations to use the accompanying premorbid system. To prevent re-learning of model parameters, the Dirichlet concentration parameters (for intrinsic and extrinsic connections) were set at high values—as if the subject had over-learned the task (see Lesion 0, Table 1).

Disabling the premorbid system and engaging the alternative system (Lesion 1). Within the premorbid (attentive) system, we introduced structural disconnections to the extrinsic connections linking state factors (Target Words, Spoken Words and Epoch) to outcomes (Proprioception, Words and Evaluation), see Fig. 1 and Table 1, Lesion 1). Now the subject is unable to use the premorbid system to differentiate between implausible and plausible hidden states because there is no evidence for any particular Target Word causing the Word outcome—and no systematic mapping between the Spoken Word and Evaluation. In light of this, the disconnected premorbid system is also unable to learn anything. In contrast, all the extrinsic connections in the alternative (inattentive) system were preserved. As in the control model (Lesion 0), re-learning of the model parameters in the alternative system was precluded with high Dirichlet concentration parameters (for both intrinsic and extrinsic connections).

Despite structural disconnections to the premorbid system, and a strong prior expectation at the higher level to use the premorbid state (as in the control model), we expect that Lesion 1 will be able to perform word repetition without any re-learning. This is because the alternative system can support the same function—albeit with lower precision. Functional re-organisation is still required, however, because the lesioned subject still needs to (i) correctly infer (at the higher level) that there has been a change in the options available (since the premorbid option has been rendered ineffective for the task at hand) and (ii) shift, at the higher level, from using the (damaged) premorbid system to the intact alternative system. This would be evidenced by a change in beliefs at the higher level of the model. If the subject fails to engage the alternative system, incorrect responses will be

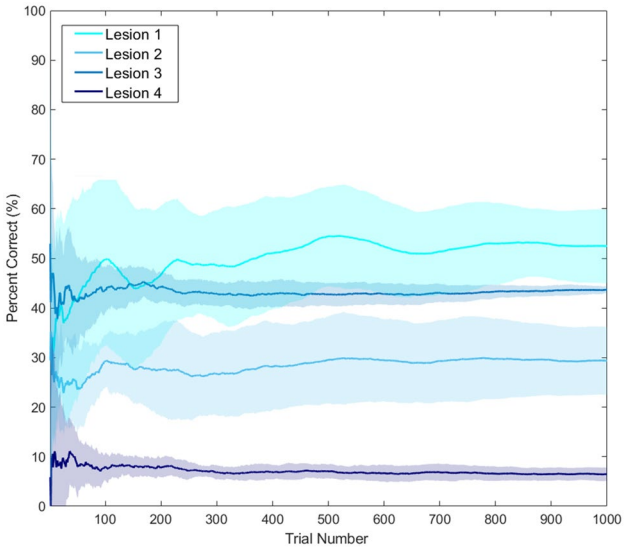

Figure 2. Behavioural performance across 1000 trials. This plot shows the percentage of correct responses for each model across 1000 trials. The x-axis is the trial number, and the y-axis is the cumulative percentage of correct responses (i.e., the percentage of trials that were correct for all trials up to the current trial number). Here, the lines report the average response, and shaded area the 95% confidence intervals, calculated across 10 simulations.

produced. In brief, any recovery depends on rapid updates to higher-level beliefs about the options available for task performance. This should happen after a few trials, when the model gets incorrect feedback, when attempting to use the premorbid system.

Long-term experience-dependent plasticity in the alternative system (Lesion 2).. The second intervention was the same as the first, but with an additional functional lesion to the intrinsic connections mediating state transitions,B^1, for the Target Word factor (see Fig. 1 and Lesion 2 in Table 1). We expected this to increase prediction errors when inferring the target word at Epoch 2.

In addition to (i) the structural lesions to extrinsic connections in the premorbid system and (ii) functional lesions to intrinsic connections which impacts on the alternative system, we also (iii) reduced the Dirichlet parameters (for intrinsic and extrinsic connections). Now the subject can accumulate Dirichlet parameters that best account for experiences (via future learning) thereby strengthening synaptic connections of model parameter B[136,46] in the alternative system. This is consistent with experience-dependent plasticity. In brief, we expected that the subject would partially recover slowly through experience-dependent plasticity within the alternative system.

Other forms of *in-silico* intrinsic lesions would have impeded the model's ability to exhibit adaptive experience-dependent plasticity. For example, introduction of mis-specified or jumbled (as opposed to imprecise) Dirichlet priors would have resulted in maladaptive learning of Dirichlet priors over the course of the simulations.

Manipulating the neuromodulatory balance between the premorbid and alternative systems (Lesion 3). To disrupt the neuromodulatory balance between the premorbid and alternative system, we induced functional lesions to the likelihood mappings for extrinsic connections to the Word, Proprioception and Evaluation outcomes, see Fig. 1 and up 3 in Table 1. Compared to Lesion 1, the premorbid system is not completely disconnected, and the alternative system is less precise. Now the subject can use both systems (albeit ineffectively) but no longer has the appropriate machinery to control which system to use. In addition, as in Lesion 0 and 1, we prevented re-learning of model parameters by using high Dirichlet concentration parameters (for intrinsic and extrinsic connections) in both systems (see Lesion 3 in Table 1).

Multifocal lesions (Lesion 4). In Lesion 4, we repeated Lesion 3 (functional lesions to extrinsic mappings in premorbid and alternative systems) and additionally lesioned the intrinsic connections as in Lesion 2, (see Table 1). These multifocal lesions were expected to further impede functional recovery relative to all prior models. Nevertheless, the system retained learning capacity because the Dirichlet concentration parameters were set to be low (see Lesion 4, Table 1).

Results of simulations

For each lesion, we evaluated word repetition performance (Fig. 2) and lesion severity (Figs. 3, 4) over 1000 trials. Lesion severity is quantified using: (i) free energy, (ii) degeneracy and (iii) redundancy. As a reminder, free energy is the complexity cost incurred in forming accurate posterior beliefs about the causes of sensations/outcomes. As policies are more probable, a priori, if they minimise the expected free energy, model evidence is

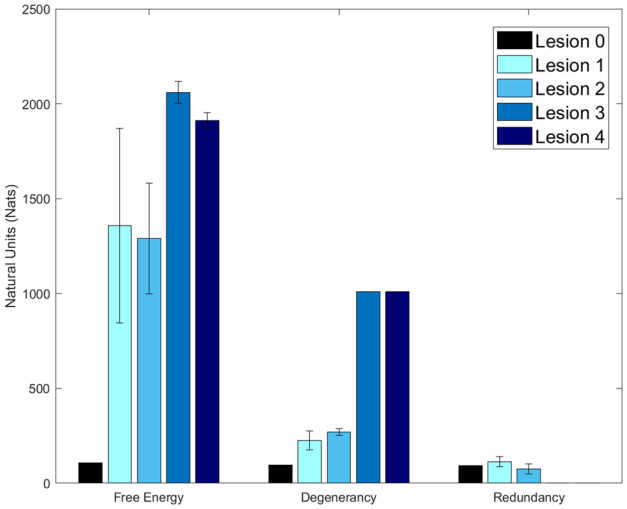

Figure 3. Free energy, degeneracy, and redundancy. This bar plot reports the total free energy, degeneracy, and redundancy for each kind of lesion across 1000 trials, with error bars calculated from 10 simulations. Free energy is the model evidence, degeneracy is the entropy of posterior beliefs about the causes of sensations and redundancy is the complexity cost incurred by forming those beliefs. The y-axis represents information, measured in natural units. Both Lesion 3 and Lesion 4 have redundancy < 30 nats—due to an inability to update posterior beliefs away from prior beliefs.

higher when free energy is lower. Degeneracy, in this formal setting, is the many-to-one mapping between hidden causes and outcomes (i.e., entropy of posterior beliefs). This has the advantage of providing flexibility in our internal explanations for sensory outcomes (keeping our options open, allowing re-learning) but can also result in less confidence (more uncertainty) in what is causing the outcomes (lowering accuracy). Finally, redundancy is the inverse of efficiency and represents the complexity cost of updating beliefs when there is multiple alternative (i.e., degenerate) causes to consider (i.e., more belief updating is required).

Control simulation (Lesion 0). In the absence of lesions, our synthetic subject was able to repeat the correct word with 100% accuracy across all 1000 trials. This indicates that the subject correctly inferred that they were operating in an 'attentive' (i.e., precise) context, relying on the premorbid system, despite being equipped with an alternative system.

Disabling the premorbid system and engaging the alternative system (Lesion 1). After a sharp drop in performance in the first 50 trials, Lesion 1 regained performance, and stabilisation to 51% [39–61% CI] by the 400th trial that persisted across the remaining trials (Fig. 2; cyan line). Overall mean performance for this model was 54% [45–60% CI].

Compared to Lesion 0 (the control model), Lesion 1 had lower (i) accuracy; (ii) model evidence (high free energy) and (iii) degeneracy (representational flexibility) because, when the premorbid system was unavailable, the alternative system (with less precise likelihood mappings) was engaged. The novel insight here is that Lesion 1 was able to shift context (to attend to the alternative system) even though the network level indicates that the premorbid system should be used. This models a higher-level reorganisation of the neural circuitry involved in repeating words that arose naturally because the premorbid system was damaged and was therefore unable to supply precise evidence to the higher level.

Long-term experience-dependent plasticity in the alternative system (Lesion 2).. Lesion 2 had impaired behavioural performance, across trials, relative to Lesion 1 (29% versus 54%) reflecting the additional lesion to intrinsic connections in the alternative system that compromised new learning. The residual learning capacity in Lesion 2 is illustrated by observing the change in Dirichlet parameters over time (Fig. 5). This shows how Lesion 2 slowly shifts towards the 'optimal' (Lesion 0) distribution, with saturation effects.

Manipulating the neuromodulatory balance between the premorbid and alternative systems (Lesion 3). In the first 50 trials, Lesion 3 performed better than Lesion 1 because the premorbid system retained some capacity (was not completely disconnected). However, unlike Lesion 1, the performance of Lesion 3 did not improve over time, stabilising at about 43%, after trial 250 (see Fig. 2). This is because (i) Lesion 3 struggled to infer which system (premorbid or alternative) was appropriate and (ii) both the premorbid and alternative systems were generating errors because precision in the extrinsic connections was low. Consequently, compared to Lesions 1 and 2, Lesion 3 had higher free energy (less model evidence) and higher degeneracy (more uncertainty) across trials. This lesion also produced lower redundancy that reflects an inability to appropriately update posterior beliefs i.e., the model can no longer estimate which causes were responsible for the outcomes.

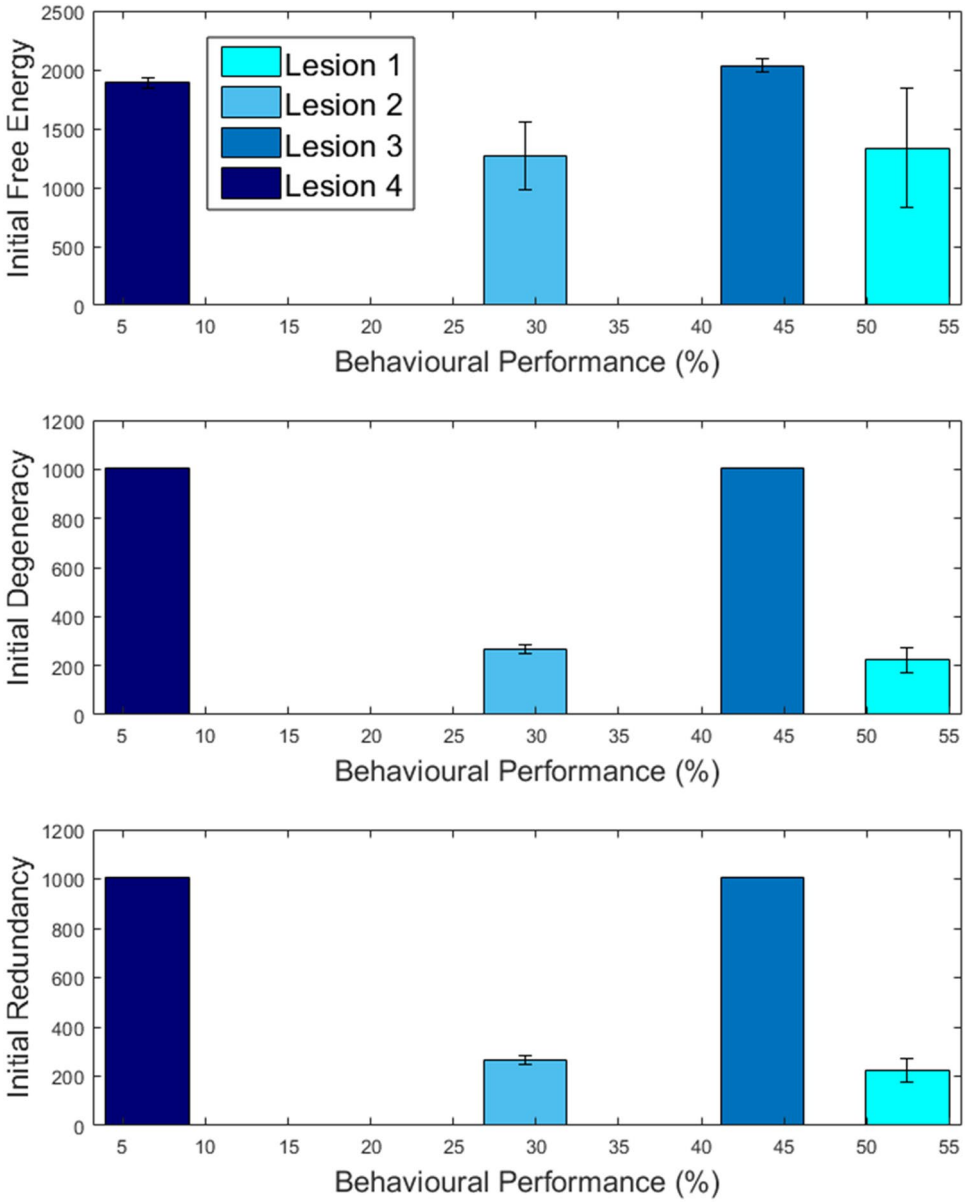

Figure 4. Lesion severity and behavioural performance. For each model, the plots report (in nats) free energy (top row), degeneracy (middle row) and redundancy (bottom row), accumulated over the first 50 trials (y axis), according to behavioural performance (percent correct responses measured after 1000 trials in x axis). The error bars are calculated from 10 simulations.

Multifocal lesions (Lesion 4). Lesion 4 performed worse than Lesion 3, across trials, demonstrating the impact of compromised learning. It's also interesting to note that Lesion 4 had less free energy than Lesion 3 (i.e. greater model evidence) because it had higher levels of redundancy and accuracy (free energy = redundancy – accuracy)[26] (Fig. 3).

Physiological predictions. Recovery has previously been associated with various physiological changes e.g., increased baseline firing frequency[23,30], synchronous neural activity[23], etc. The *in-silico* lesions reflect these changes in neuronal functionality due to missing/interrupted circuitry (via degeneracy) and alterations within surviving structures (via peri-lesional activity) that affect the belief updating process.

The lesioned models demonstrate reduced baseline evoked response frequency and a shifted inhibitory evoked response, relative to control (Fig. 6). Lesion 1 has similar evoked response magnitude to control, but a decreased presynaptic excitatory potential across all trials (Fig. 6; second row)—similar trends have been observed in aphasic subjects during picture naming tasks[57]. Additionally, the initial trials exhibit an attenuated evoked response, but later trials expose a steady increase in evoked responses[57]. Lesion 2 has an attenuated, but

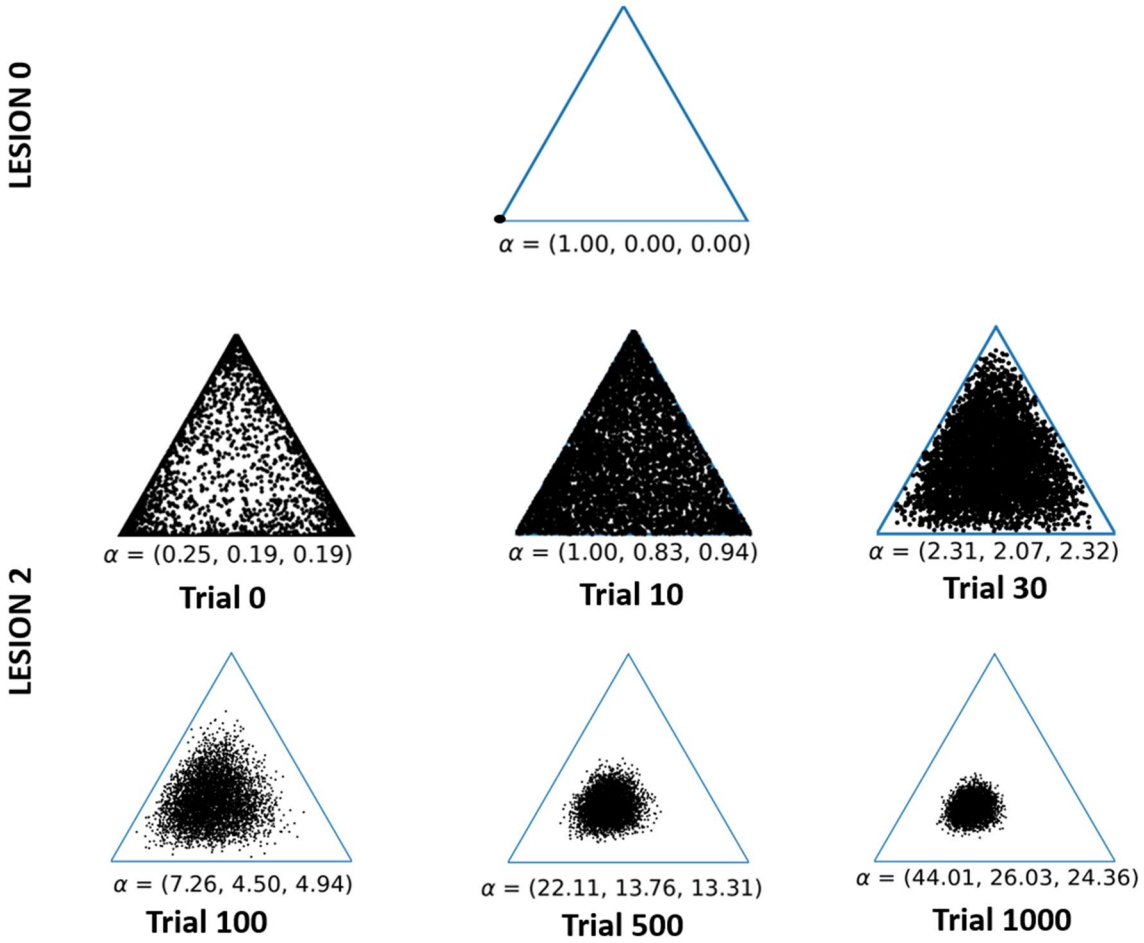

Figure 5. Synaptic strength changes via experience-dependent plasticity. The figure illustrates the Dirichlet distribution in a 3-dimensional coordinate space, i.e., 2-simplex for a particular simulation. The concentration of dots in one corner reflect precise beliefs; and scattered dots denote imprecise beliefs. Each dot represents a single sample from the Dirichlet distribution (determined by the alpha parameters), and each plot displays 5000 samples. For visual clarity, we have focused on the changes in the Target Word transitions for how the word 'red' can potentially transition to either 'red', 'blue' or 'table'. Similar distributional shifts (i.e., learning) apply to other Target Word transitions. The plot for the control model illustrates that, with a precise transition mapping, the distribution is highly concentrated at one corner of the plot (to the extent that the points look like a single dot). On the second row, a series of panels illustrate the learning process for Lesion 2 over six time points (trials 0, 10, 30, 100, 500 and 1000). Immediately following the lesion (trial 0) a scattered distribution is evident, despite a higher concentration of points at all corners of the triangle. However, as the lesioned model learns (between trials 0 and 1000), the distribution converges to the corner associated with the high alpha—more closely resembling the control.

noticeable, presynaptic inhibitory potential (Fig. 6; third row) and muted overall excitatory potential. Lesion 3 has the highest decrease in evoked response, compared to other models, alongside an decrease in postsynaptic inhibitory potential (Fig. 6; fourth row)[58]. Muted evoked response can be observed for Lesion 4 (Fig. 6; fifth row).

This simulated electrophysiology serves three key purposes. Firstly, it lends a construct validity to the generative model for word repetition (and the inferential process); these simulations are congruent with real electrophysiological measurements from humans[57–59] and other primates[30,31]. Second, it offers an intuition about the computational processes that might underpin these responses, and how they change when particular processes are perturbed. Finally, we can derive quantitative predictions about measured electrophysiology from human subjects and use them to quantify recovery mechanisms. Our simulations suggest that different perturbations lead to different neuronal responses, so it should be possible—in principle—to use electrophysiological data to discriminate between different types of impairment. However, further work is required to anatomically ground our word repetition models.

Discussion

In this study, we developed a generative model that is equipped with two distinct systems for independently performing a word repetition task. One system has high sensory precision allowing it to reproduce attentive and accurate behaviour. The other system has lower sensory precision that generated less accurate responses.

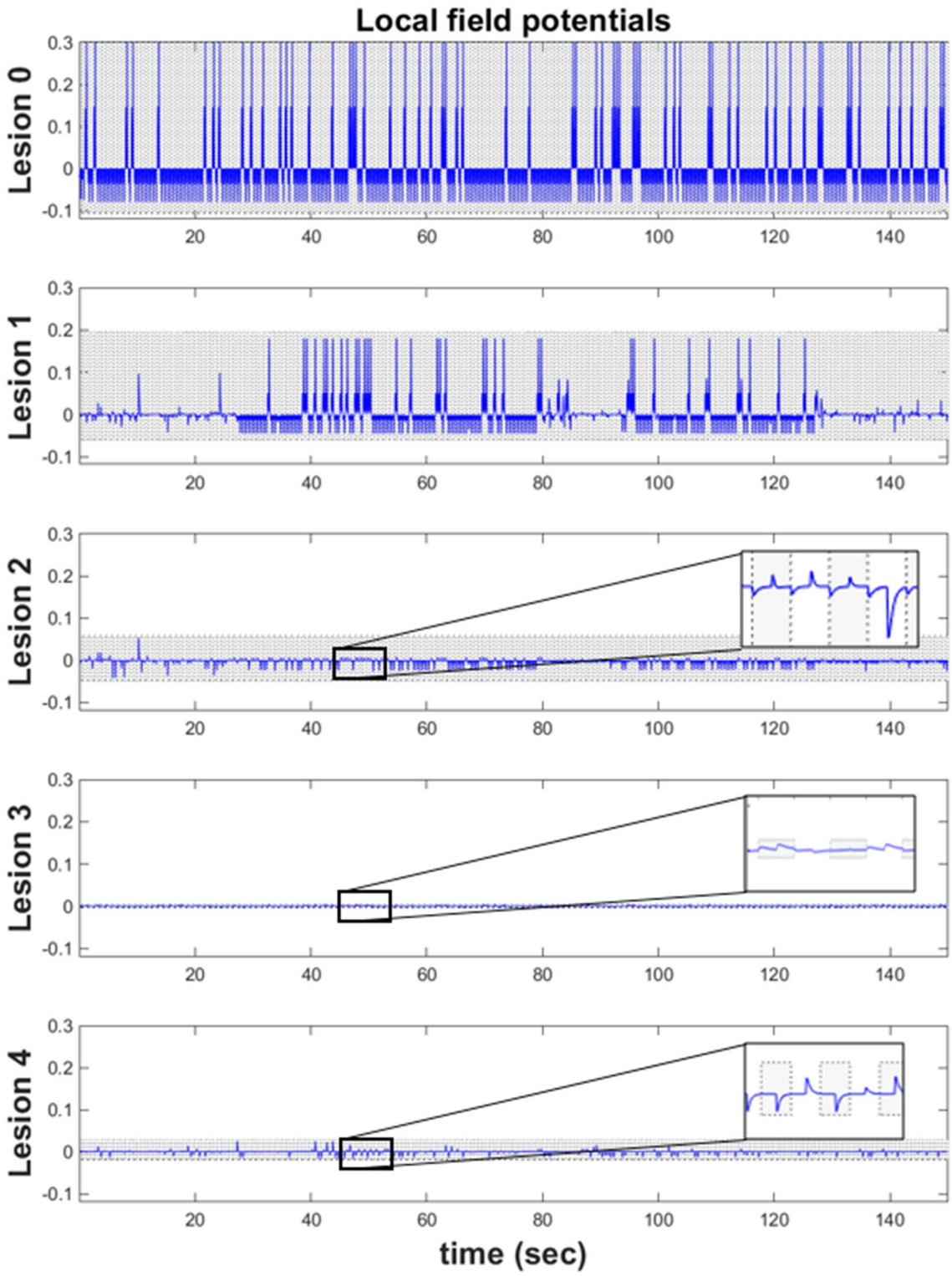

Figure 6. (Simulated local field potentials) These plots show the temporal changed in simulated local field potentials for the Target Word 'table' neuronal population across the first 300 trials, for a particular simulation. The blue line represents the trajectory of evoked responses for the Target Word 'table' over some arbitrary unites (y-axis). The simulated lesion models are presented in the following order: Lesion 0, Lesion 1, Lesion 2, Lesion 3 and Lesion 4.

The model selects, a priori, the most precise system available to perform the task, therefore, in the absence of any lesions (i.e., pre-morbidly), the model chooses to use the more precise system. Anecdotally, it attends to the task at hand.

We lesioned this model in four different ways and simulated task performance over 1000 trials to compare lesion severity and the time course of recovery. In Lesion 1, we demonstrate that, when the premorbid system is unavailable (after extensively disconnecting extrinsic connections), the alternative system (which is now the most precise) is engaged almost immediately after the synthetic recognises something has changed. The subject is then able to sustain performance with ~ 51% accuracy after approximately 400 trials. In Lesion 2, we demonstrate the effect of learning in Lesion 1 by showing that the same level of performance is precluded when the capacity to relearn is impaired (by lowering precision in the intrinsic connections). In Lesion 3, we show that incomplete damage to the extrinsic connections in both systems resulted in declining performance with no evidence of recovery—and worse performance than complete damage to one system (i.e., Lesion 1). Finally, the performance of Lesion 3 was even worse when it had reduced capacity to relearn (i.e., Lesion 4). From this we infer that the preserved intrinsic connections in Lesion 3 prevented further decline (i.e., like that observed in Lesion 4).

Our work provides a further step towards integrating the theoretical construct of degeneracy[26] and clinical patient behavioural data—through in-silico lesions of active inference models, i.e., computational neuropsychology. It offers intuition for what type of functional recovery mechanisms might underwrite particular behavioural profiles, in the context of a word repetition task. This understanding is essential for interpreting changes, in free energy or degeneracy, for word repetition models inverted using patient data[60]. Furthermore, the approach is generic and can be applied to other paradigms that investigate language; e.g., picture naming[61] or speech[62].

Future work could investigate the effect of changing the precision of other extrinsic connections. For example, an extrinsic disconnection that renders the Target Word and Spoken Word conditionally independent of each other might show similar performance deficits to those from intrinsic lesion to the Target Word factor. Similarly, recovery via degeneracy could be investigated by disconnecting specific sub-structures of the extrinsic pathways. This would speak to the possibility that biological lesions trigger multiple recovery mechanisms, individually or together, to engender resistance to functional loss. Future studies could also investigate how performance changes when functional reorganisation and re-learning are required to establish a system that was not readily available, (e.g., when the alternative system had extremely imprecise priors before the intervention). In this case, we would expect a sustained drop in performance, imprecise beliefs about the nature of recovery, and longer recovery times, as the model learns to use (i.e., reconfigure via structure learning) an alternative system. Likewise, testing how recovery changes with the degree of asymmetry between the premorbid and alternative models (i.e., imbalanced resource allocation between attentive and inattentive context) may reveal a limited capacity to call upon intact structures and may feature a permanent performance deficit, as observed in chronic patients who fail to fully recover after brain damage.

The impact of other factors beyond lesion type (e.g., age, training intensity) on recovery mechanism[20]—would also provide further avenues for interesting future work. For example, the generative model could be equipped with particular priors that enable it to mimic these features—or perhaps it could be given a particular set of experiences (i.e., exposure trials) from which it learns particular parameters. A better understanding of these functional recovery mechanisms—and potential reorganisation processes—may even help target therapeutic strategies after brain insult and improve the effectiveness of rehabilitation[63–65].

Conclusion

Our simulations reveal that recovery depends on the availability of computational and structural resources. While the model could develop resistance to insult, the effects of damage could not be overcome beyond a certain point, leading to persistent impairments. The same model was used to make physiological predictions, by simulating neuronally plausible Bayesian belief updating. The simulated lesions resulted in varied decline in baseline evoked responses. These quantitative predictions indicate the potential of future developments to investigate the neurophysiology of functional recovery and could allow us to infer likely damage based on a patient's electrophysiological responses and/or functional recovery profile.

Code availability

The generative model in these kind of simulations changes from application to application; however, the belief updates are generic and can be implemented using standard routines (here spm_MDP_VB_X.m). These routines are available as Matlab code in the SPM academic software: http://www.fil.ion.ucl.ac.uk/spm/. The code to replicate this particular generative model and accompanying stimulation data is available: https://github.com/ucbtns/frec.

Received: 30 June 2020; Accepted: 22 March 2021
Published online: 02 April 2021

References

1. Connolly, A. M., Dodson, W. E., Prensky, A. L. & Rust, R. S. Course and outcome of acute cerebellar ataxia. *Lancet Neurol.* **35**, 673–679. https://doi.org/10.1002/ana.410350607 (1994).
2. Langhorne, P., Coupar, F. & Pollock, A. Motor recovery after stroke: A systematic review. *Lancet Neurol.* **8**, 741–754 (2009).
3. Bultmann, U. *et al.* Functional recovery and rehabilitation of postural impairment and gait ataxia in patients with acute cerebellar stroke. *Gait Posture* **39**, 563–569. https://doi.org/10.1016/j.gaitpost.2013.09.011 (2014).
4. Seghier, M. L. *et al.* Visual recovery after perinatal stroke evidenced by functional and diffusion MRI: Case report. *BMC Neurol.* **5**, 17 (2005).
5. Guzzetta, A. *et al.* Plasticity of the visual system after early brain damage. *Dev. Med. Child Neurol.* **52**, 891–900 (2010).

6. Hope, T. M. H. *et al.* Recovery after stroke: Not so proportional after all?. *Brain* **142**, 15–22. https://doi.org/10.1093/brain/awy302 (2019).
7. Saur, D. *et al.* Dynamics of language reorganization after stroke. *Brain* **129**, 1371–1384 (2006).
8. Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L. & Sporns, O. Modeling the impact of lesions in the human brain. *PLoS Comput. Biol.* **5**, e1000408. https://doi.org/10.1371/journal.pcbi.1000408 (2009).
9. Irle, E. Lesion size and recovery of function: Some new perspectives. *PLoS Comput. Biol.* **12**, 307–320. https://doi.org/10.1016/0165-0173(87)90003-8 (1987).
10. Chen, C.-L., Tang, F.-T., Chen, H.-C., Chung, C.-Y. & Wong, M.-K. Brain lesion size and location: Effects on motor recovery and functional outcome in stroke patients. *Arch. Phys. Med. Rehabil.* **81**, 447–452. https://doi.org/10.1053/mr.2000.3837 (2000).
11. Warburton, E., Price, C. J., Swinburn, K. & Wise, R. J. S. Mechanisms of recovery from aphasia: Evidence from positron emission tomography studies. *J. Neurol. Neurosurg. Psychiatry* **66**, 155–161. https://doi.org/10.1136/jnnp.66.2.155 (1999).
12. Welbourne, S. R., Woollams, A. M., Crisp, J. & Lambon-Ralph, M. A. The role of plasticity-related functional reorganization in the explanation of central dyslexias. *Cogn. Neuropsychol.* **28**, 65–101 (2011).
13. Seghier, M. L. *et al.* Reading without the left ventral occipito-temporal cortex. *Neuropsychologia* **50**, 3621–3635. https://doi.org/10.1016/j.neuropsychologia.2012.09.030 (2012).
14. Price, C. J. & Friston, K. J. Degeneracy and cognitive anatomy. *Trends Cogn. Sci.* **6**, 416–421 (2002).
15. Tononi, G., Sporns, O. & Edelman, G. M. Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci.* **96**, 3257–3262. https://doi.org/10.1073/pnas.96.6.3257 (1999).
16. Nudo, R. J. Adaptive plasticity in motor cortex: Implications for rehabilitation after brain injury. *J. Rehabil. Med.* **35**, 7–10. https://doi.org/10.1080/16501960310010070 (2003).
17. Fu, M. & Zuo, Y. Experience-dependent structural plasticity in the cortex. *Trends Neurosci.* **34**, 177–187 (2011).
18. Lövdén, M., Wenger, E., Mårtensson, J., Lindenberger, U. & Bäckman, L. Structural brain plasticity in adult learning and development. *Neurosci. Biobehav. Rev.* **37**, 2296–2310 (2013).
19. Nudo, R. Recovery after brain injury: mechanisms and principles. *Front. Hum. Neurosci.* https://doi.org/10.3389/fnhum.2013.00887 (2013).
20. Kleim, J. A. & Jones, T. A. Principles of experience-dependent neural plasticity: Implications for rehabilitation after brain damage. *J. Speech Lang. Hear. Res.* **51**, S225-239. https://doi.org/10.1044/1092-4388(2008/018) (2008).
21. Cooper, S. J. & Donald, O. Hebb's synapse and learning rule: A history and commentary. *Neurosci. Biobehav. Rev.* **28**, 851–874 (2005).
22. Hope, T. M. H. *et al.* Right hemisphere structural adaptation and changing language skills years after left hemisphere stroke. *NeuroImage Clin.* **140**, 1718–1728 (2017).
23. Carmichael, S. T. Plasticity of cortical projections after stroke. *Neuroscientist* **9**, 64–75 (2003).
24. Ueno, T., Saito, S., Rogers, T. T. & Lambon-Ralph, M. A. Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron* **72**, 385–396. https://doi.org/10.1016/j.neuron.2011.09.013 (2011).
25. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active inference: A process theory. *Neural Comput.* **29**, 1–49. https://doi.org/10.1162/NECO_a_00912 (2017).
26. Sajid, N., Parr, T., Hope, T. M., Price, C. J. & Friston, K. J. Degeneracy and redundancy in active inference. *Cereb. Cortex* https://doi.org/10.1093/cercor/bhaa148 (2020).
27. Sajid, N., Parr, T., Gajardo-Vidal, A., Price, C. J. & Friston, K. J. Paradoxical lesions, plasticity and active inference. *Brain Commun.* https://doi.org/10.1093/braincomms/fcaa164 (2020).
28. Hope, T. M. H. *et al.* Dissecting the functional anatomy of auditory word repetition. *Front. Hum. Neurosci.* **8**, 246–246. https://doi.org/10.3389/fnhum.2014.00246 (2014).
29. Hickok, G. The architecture of speech production and the role of the phoneme in speech processing. *Lang. Cogn. Process* **29**, 2–20. https://doi.org/10.1080/01690965.2013.834370 (2014).
30. Schiene, K. *et al.* Neuronal hyperexcitability and reduction of GABAA-receptor expression in the surround of cerebral photothrombosis. *J. Cereb. Blood Flow Metab.* **16**, 906–914 (1996).
31. Luhmann, H. J., Mudrick-Donnon, L. A., Mittmann, T. & Heinemann, U. Ischaemia-induced long-term hyperexcitability in rat neocortex. *Eur. J. Neurosci.* **7**, 180–191 (1995).
32. Neumann-Haefelin, T., Hagemann, G. & Witte, O. W. Cellular correlates of neuronal hyperexcitability in the vicinity of photochemically induced cortical infarcts in rats in vitro. *Neurosci. Lett.* **193**, 101–104 (1995).
33. Friston, K. A free energy principle for a particular physics. http://arxiv.org/abs/1906.10184 (2019).
34. Sajid, N., Ball, P. J. & Friston, K. J. Active inference: demystified and compared. http://arxiv.org/abs/1909.10863 (2019).
35. Parr, T. & Friston, K. J. Generalised free energy and active inference: can the future cause the past?. *bioRxiv* https://doi.org/10.1101/304782 (2018).
36. Da Costa, L. *et al.* Active inference on discrete state-spaces: A synthesis. http://arxiv.org/abs/2001.07203 (2020).
37. Friston, K. J., Rosch, R., Parr, T., Price, C. & Bowman, H. Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* **77**, 388–402. https://doi.org/10.1016/j.neubiorev.2017.04.009 (2017).
38. Parr, T., Markovic, D., Kiebel, S. J. & Friston, K. J. Neuronal message passing using mean-field, bethe, and marginal approximations. *Sci. Rep.* **9**, 1889. https://doi.org/10.1038/s41598-018-38246-3 (2019).
39. Friston, K. J., Parr, T. & de Vries, B. The graphical brain: Belief propagation and active inference. *Netw. Neurosci.* **1**, 381–414. https://doi.org/10.1162/NETN_a_00018 (2017).
40. Parr, T., Rikhye, R. V., Halassa, M. M. & Friston, K. J. Prefrontal computation as active inference. *Cereb. Cortex* **30**, 682–695 (2019).
41. Friston, K. J. *et al.* Active inference, curiosity and insight. *Neural Comput.* **29**, 2633–2683. https://doi.org/10.1162/neco_a_00999 (2017).
42. Parr, T. *The Computational Neurology of Active Vision* (University College London, 2019).
43. Beal, M. J. Variational Algorithms for Approximate Bayesian Inference. *PhD. Thesis, University College London* (2003).
44. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
45. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York, 1949).
46. Friston, K. *et al.* Active inference and learning. *Neurosci. Biobehav. Rev.* **68**, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022 (2016).
47. Moran, R. J. *et al.* Free energy, precision and learning: The role of cholinergic neuromodulation. *J. Neurosci.* **33**, 8227–8236. https://doi.org/10.1523/jneurosci.4255-12.2013 (2013).
48. Nozari, N. & Dell, G. S. How damaged brains repeat words: A computational approach. *Brain Lang.* **126**, 327–337. https://doi.org/10.1016/j.bandl.2013.07.005 (2013).
49. Moritz-Gasser, S. & Duffau, H. The anatomo-functional connectivity of word repetition: insights provided by awake brain tumor surgery. *Front. Hum. Neurosci.* **7**, 405 (2013).
50. Parr, T. & Friston, K. J. Uncertainty, epistemics and active inference. *J. R. Soc. Interface* **14**, 20170376. https://doi.org/10.1098/rsif.2017.0376 (2017).
51. Brown, H., Friston, K. J. & Bestmann, S. Active inference, attention, and motor preparation. *Front. Psychol.* **2**, 1–10. https://doi.org/10.3389/fpsyg.2011.00218 (2011).

52. Chang, Y.-N. & Lambon-Ralph, M. A. A unified neurocomputational bilateral pathway model of spoken language production in healthy participants and recovery in post-stroke aphasia. *bioRxiv* **6**, 337 (2020).
53. Tourville, J. A. & Guenther, F. H. The DIVA model: A neural theory of speech acquisition and production. *Lang. Cogn. Process* **26**, 952–981. https://doi.org/10.1080/01690960903498424 (2011).
54. Guenther, F. H. & Vladusich, T. A neural theory of speech acquisition and production. *J. Neurolinguistics* **25**, 408–422. https://doi.org/10.1016/j.jneuroling.2009.08.006 (2012).
55. Houde, J. & Nagarajan, S. Speech production as state feedback control. *Front. Hum. Neurosci.* https://doi.org/10.3389/fnhum.2011.00082 (2011).
56. Perrinet, L. U., Adams, R. A. & Friston, K. J. Active inference, eye movements and oculomotor delays. *Biol. Cybern.* **108**, 777–801 (2014).
57. Laganaro, M., Morand, S., Schwitter, V., Zimmermann, C. & Schnider, A. Normalisation and increase of abnormal ERP patterns accompany recovery from aphasia in the post-acute stage. *Neuropsychologia* **46**, 2265–2273. https://doi.org/10.1016/j.neuropsychologia.2008.02.013 (2008).
58. Kotz, S. A. & Friederici, A. D. Electrophysiology of normal and pathological language processing. *J. Neurolinguistics* **16**, 43–58 (2003).
59. Pei, X. *et al.* Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage* **54**, 2960–2972. https://doi.org/10.1016/j.neuroimage.2010.10.029 (2011).
60. Schwartenbeck, P. & Friston, K. Computational phenotyping in psychiatry: A worked example. *eNeuro* https://doi.org/10.1523/ENEURO.0049-16.2016 (2016).
61. Sajid, N., Friston, K. J., Ekert, J. O., Price, C. J. & Green, D. Neuromodulatory control and language recovery in bilingual aphasia: An active inference approach. *Behav. Sci.* **10**, 161 (2020).
62. Friston, K. J. *et al.* Active listening. *Hearing Res.* **399**, 107998 (2020).
63. Berthier, M. L. & Pulvermuller, F. Neuroscience insights improve neurorehabilitation of poststroke aphasia. *Nat. Rev. Neurol.* **7**, 86–97. https://doi.org/10.1038/nrneurol.2010.201 (2011).
64. Chua, K. S. & Kong, K.-H. Functional outcome in brain stem stroke patients after rehabilitation. *Arch. Phys. Med. Rehabil.* **77**, 194–197 (1996).
65. Taub, E., Uswatte, G. & Elbert, T. New treatments in neurorehabilitation founded on basic research. *Nat. Rev. Neurosci.* **3**, 228–236. https://doi.org/10.1038/nrn754 (2002).

Author contributions

Conceptualization, N.S., C.J.F., & K.J.F; Defining the generative model & formal analysis, N.S.; Software, N.S, and K.J.F; Writing—original draft, N.S.; Writing—review & editing, E.H., T.H., Z.F., C.J.F., and K.J.F. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by Medical Research Council (MR/S502522/1, NS; MR/M023672/1, CJP), Wellcome Trust (Ref: 203147/Z/16/Z and 205103/Z/16/Z, CJP and KJF; WT091681MA, EH), and Stroke Association (TSA_PDF_2017/02, TMH).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.