This is the accepted manuscript made available via CHORUS. The article has been published as:

Violations of the fluctuation-dissipation theorem reveal distinct nonequilibrium dynamics of brain states

Gustavo Deco, Christopher W. Lynn, Yonatan Sanz Perl, and Morten L. Kringelbach

Phys. Rev. E **108**, 064410 — Published 26 December 2023

DOI: 10.1103/PhysRevE.108.064410

Violations of the fluctuation-dissipation theorem reveal distinct non-equilibrium dynamics of brain states

Gustavo Deco[1,2], Christopher W. Lynn[3,4], Yonatan Sanz Perl[1,5,6], and Morten L. Kringelbach[7,8,9]

1. Center for Brain and Cognition, Computational Neuroscience Group, Universitat Pompeu Fabra, Roc Boronat 138, Barcelona, 08018, Spain

2. Institució Catalana de la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, 08010, Spain

3. Initiative for the Theoretical Sciences, Graduate Center, City University of New York, New York, NY 10016

4. Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544

5. Department of Physics, University of Buenos Aires, Buenos Aires, Argentina

6. Paris Brain Institute (ICM), Paris, France

7. Centre for Eudaimonia and Human Flourishing, Linacre College, University of Oxford, Oxford, UK.

8. Department of Psychiatry, University of Oxford, Oxford, UK

9. Center for Music in the Brain, Department of Clinical Medicine, Aarhus University, Aarhus, DK.

The brain is a non-equilibrium system whose dynamics change in different brain states, such as wakefulness and deep sleep. Thermodynamics provides the tools for revealing these non-equilibrium dynamics. We used violations of the fluctuation-dissipation theorem to describe the hierarchy of non-equilibrium dynamics associated with different brain states. Together with a whole-brain model fitted to empirical human neuroimaging data, and deriving the appropriate analytical expressions, we were able to capture the deviation from equilibrium in different brain states that arises from asymmetric interactions and hierarchical organisation.

I. INTRODUCTION

A key unsolved question in neuroscience is how different brain states such as wakefulness and deep sleep are associated with different levels of non-equilibrium brain dynamics [1, 2]. Given that non-equilibrium is intrinsically linked to irreversibility and entropy production [3], thermodynamics has shown great promise for characterising the hierarchical dynamics of brain states over time[4]. Here, we present a new framework that uses the violations of the fluctuation-dissipation theorem (FDT) to describe the non-equilibrium dynamics associated with a given brain state. Specifically, we show that perturbing a generative whole-brain model can be used to quantify the level of non-equilibrium through violations of the FDT in empirical neuroimaging data from human participants during different brain states (wakefulness, deep sleep and cognitive tasks).

This provides a new way to quantify the system-wide response of the brain to targeted perturbations and a solid theoretical framework needed for a highly influential set of neuroscience experiments that used direct perturbations of the brain in different states to measure the fluctuations and dissipation of brain activity after perturbations [2, 5, 6]. In order to assess the brain-wide spatiotemporal propagation of external stimulation the authors introduced the Perturbational Complexity Index (PCI), which measures the amount of information contained in the amplitude of the average perturbation-evoked responses by calculating the Lempel-Ziv complexity of the binary matrix describing the statistically significant sources, in space and time, of the EEG signals [2]. PCI has been successfully used for separation of brain states in healthy subjects during wakefulness, dreaming, sleep, under different levels of anaesthesia and in coma [2, 5, 6]. However, here we propose that these results are best understood in terms of violations of the FDT.

The present framework is related to recent advances in using thermodynamics to describe whole-brain dynamics [1, 4, 7, 8], which has started to identify important changes in the hierarchical organisation and orchestration in different brain states. Specifically, by quantifying the arrow of time, one can directly measure the 'breaking of the detailed balance' in non-equilibrium brain systems and thereby assess the asymmetry in the flow of information. In the brain, a useful definition of hierarchy is the asymmetrical relationship between feed-forward and feed-backward interactions between brain regions. As such, a flat hierarchy is symmetric (resulting in an equilibrium system with reversible dynamics), while a hierarchical system has asymmetric interactions (resulting in irreversible dynamics that break detailed balance and diverge from equilibrium).

Here, we move beyond these model-free measures of irreversibility to create a model-based FDT framework, offering a complementary thermodynamic perspective for describing whole-brain dynamics. The FDT framework naturally uses perturbations to quantify the degree of non-equilibrium and consequently the hierarchical organisation of brain state dynamics. Furthermore, our novel framework allows us for the first time to estimate the asymmetry in the generators of the brain dynamics. This is important since asymmetry cannot be established from conventional measures such as functional connectivity, which are symmetric by definition and do not provide any insights into the generative mechanisms underlying brain dynamics.

Overall, building on previous empirical research quantifying brain states measured with electroencephalography following transcranial magnetic stimulation (TMS) [2, 5, 6], the proposed FDT framework quantifies a brain state based on empirical functional magnetic resonance imaging (fMRI) data without any need for empirical stimulation. Instead, the FDT framework creates a whole-brain model of this empirical data, which can then exhaustively stimulated *in silico* in the whole-brain model and thus provide insight into the generative

mechanisms of brain dynamics, allowing for clear differentiation of brain states.

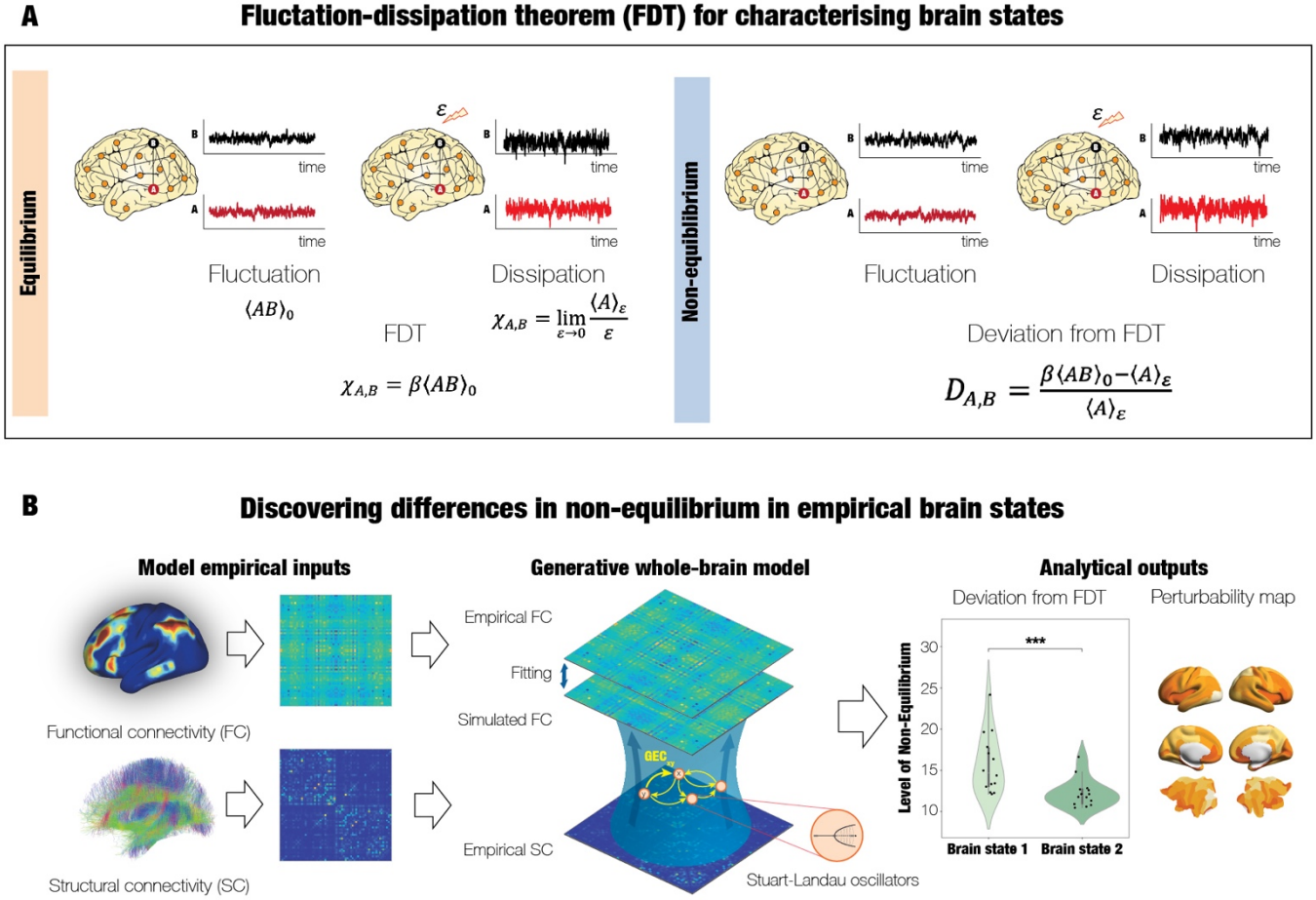


FIG 1. Fluctuation-dissipation theorem (FDT) used on empirical neuroimaging data. (a) As can be seen from the general framework of FDT in equilibrium (left panel) and non-equilibrium (right panel), this can be used to characterise different brain states. Specifically, the level of non-equilibrium can be captured as the deviation of FDT and can subsequently be used to describe the orchestration and changes in hierarchy. (b) Combining FDT with a whole-brain model (linking anatomical connectivity and functional brain connectivity) fitted to empirical neuroimaging data can precisely describe the overall deviation from FDT as well as the perturbability maps for different brain states.

II. VIOLATIONS OF FLUCTUATION-DISSIPATION THEOREM

In this section we describe how to measure the violations of FDT to quantify non-equilibrium dynamics.

A. Fluctuation-dissipation theorem

To investigate the violations of FDT, we follow Onsager, who proposed a simple derivation using his regression principle [9-11]. This principle holds that when a system begins at an initial equilibrium state and is driven by a weak external perturbation to a final equilibrium state, the evolution of the system can be treated as a spontaneous equilibrium fluctuation. Specifically, let us assume that a weak external perturbation ε is coupled to an observable B at time $t = 0$. Applying Onsager's regression principle, one can derive an expression for the difference between $\langle A(t) \rangle_\varepsilon$ (the expectation value of a second observable A after the perturbation is applied in B) and $\langle A(t) \rangle_0$ (the expectation value in the unperturbed state), which is given namely by:

$$\langle A(t) \rangle_\varepsilon - \langle A(t) \rangle_0 = \beta \varepsilon [\langle A(t) B(t) \rangle_0 - \langle A(t) B(0) \rangle_0], \quad (1)$$

where β is the inverse temperature from equilibrium thermodynamics. The time-dependent susceptibility is then given by:

$$\chi_{A,B}(t) = \frac{\partial \langle A(t) \rangle}{\partial \varepsilon} = \lim_{\varepsilon \to 0} \frac{\langle A(t) \rangle_\varepsilon - \langle A(t) \rangle_0}{\varepsilon} = \beta [\langle A(t) B(t) \rangle_0 - \langle A(t) B(0) \rangle_0], \quad (2)$$

The static form of the FDT is easily obtained by taking the limit $t \to \infty$. In this case,

$$\chi_{A,B} = \beta [\langle AB \rangle_0 - \langle A \rangle_0 \langle B \rangle_0], \quad (3)$$

since correlations factorise for infinitely separated times (see Appendix B for a detailed derivation for spin systems). Thus, in equilibrium, we arrive at a correspondence between the response of a system to perturbation (on the left-hand side) and its unperturbed correlations (on the right-hand side).

B. Violations of FDT in non-equilibrium

To characterise the level of *non-equilibrium,* we can examine the normalised deviation of the system from the FDT:

$$D_{A,B} = \frac{\beta\langle AB\rangle_0 - \chi_{A,B}}{\chi_{A,B}}, \quad (4)$$

which is obtained (without loss of generality) by defining the unperturbed state such that the mean values of the observables are set to zero; i.e., $\langle A\rangle_0 = \langle B\rangle_0 = 0$. In the numerator, the first term, $\beta\langle AB\rangle_0$, corresponds to unperturbed fluctuations, while the second term, $\chi_{A,B} = \langle A\rangle_\varepsilon/\varepsilon$, corresponds to the response to a small perturbation ε. The total deviation D can be obtained by averaging $D_{A,B}$ over all observables A and all perturbation sites B. Hence, the degree of violation of the FDT, quantified by D, measures the divergence of the system from equilibrium. In turn, we hypothesize that these violations of the FDT will result from asymmetries in the interactions within a system, which can change from one brain state (e.g., resting versus performing a cognitive task) to another.

III. MODEL-BASED FDT OF WHOLE-BRAIN NEUROIMAGING DATA

To test this hypothesis, we investigate the spatiotemporal dynamics underlying radically different brain states using empirical human neuroimaging data recorded using functional Magnetic Resonance Imaging (fMRI).

A. Theoretical framework

Figure 1 summarizes the main paradigm. In order to estimate the total deviation from the FDT for each participant in a given brain state, we first construct a whole-brain model fitting the corresponding functional neuroimaging data. This allows us to derive analytical expressions for the correlations between all brain regions under spontaneous fluctuations and the effect of a perturbations in one brain region on the average activities of all other regions across the brain. This whole-brain model-based analytical expression can be used to derive the total deviation from the FDT. In Equation 4, D can be estimated after exhaustively perturbing all brain regions B and observing the corresponding effects on all brain regions A.

To investigate the system-wide response of neural activity to targeted perturbations, we require a model of whole-brain dynamics. Here we build on the rich literature over the last ten years linking anatomical structural connectivity and functional dynamics[12-15]. The anatomical structural connectivity (SC) can be determined *in vivo* using diffusion MRI (dMRI) in conjunction with probabilistic tractography, leading to what is commonly known as the structural connectome. The whole-brain model of neural activity strikes a compromise between complexity and realism by using the physical wiring between brain regions (reflected in SC) to reproduce the empirically-measured whole-brain dynamics recorded using fMRI [15]. Such whole-brain models have had widespread success in explaining the patterns of spontaneous correlations between brain regions, forming the so-called resting-state networks [16-21].

B. Whole-brain model

Here, we modelled the local dynamics of each brain region as a Stuart-Landau oscillator (i.e., as the normal form of a supercritical Hopf bifurcation), the standard model for examining the shift from noisy to oscillatory dynamics [22]. Whole-brain Hopf models have been able to replicate key aspects of brain dynamics observed in electrophysiology [23, 24], magnetoencephalography [25] and fMRI [26, 27]. Specifically, given a parcellation of N regions, the whole-brain dynamics can be expressed by coupling the local dynamics of N Stuart-Landau oscillators via the connectivity matrix $\boldsymbol{C}$, which is defined by

$$\frac{dz_j}{dt} = (a_j + \mathrm{i}\omega_j)z_j - |z_j|^2 z_j + \sum_{k=1}^{N} C_{jk}(z_k - z_j) + \eta_j, \quad (5)$$

where the complex variable z_j denotes the state ($z_j = x_j + \mathrm{i}y_j$) of region j, η_j is additive uncorrelated Gaussian noise with variance σ^2 (for all j), ω_j is the intrinsic node frequency, and a_j is the node's bifurcation parameter. The intrinsic frequencies ω_j (which lie in the 0.008–0.08Hz band) were estimated from the data as the averaged peak frequencies of the narrowband blood-oxygen-level-dependent (BOLD) signals of the different brain regions. For $a_j > 0$, the local dynamics settle into a stable limit cycle, producing self-sustained oscillations with frequency $\omega_j/(2\pi)$. For $a_j < 0$, the local dynamics present a stable spiral point, producing damped or noisy oscillations in the absence or presence of noise, respectively. The fMRI signals were modelled by the real part of the state variables; i.e., $x_j = \mathrm{Real}(z_j)$.

It has been shown that the best working point for fitting whole-brain neuroimaging dynamics is at the brink of the bifurcation, i.e. with a_j slightly negative but very near to zero (usually $a_j = -0.02$) [28]. In other words, the dynamics operates near criticality, consistently with several works from the statistical [29, 30] and dynamical system perspective [31, 32]. In particular, from a statistical perspective, it has been described that different sleep stages can be characterised by different critical exponent, but still operating within a critical regime [33]. From a dynamical system view it has also demonstrated that whole-models operating close to the critical point are the best option to capture the dynamics of wakefulness and sleep stages represented by differences in bifurcation parameters [34, 35]. This proximity to criticality is crucial, because it allows a linearization of the dynamics, which, in turn, permits an analytical solution for the functional connectivity matrix **FC,** given by the Pearson correlations between all pairs of brain regions. We can estimate the functional correlations of the whole-brain network using a linear noise approximation (LNA). Hence, the dynamical system of N nodes (Equation 5) can be re-written in vector form as:

$$\frac{d\boldsymbol{z}}{dt} = (\boldsymbol{a} - \boldsymbol{S} + \mathrm{i}\boldsymbol{\omega})\odot\boldsymbol{z} - (\boldsymbol{z}\odot\bar{\boldsymbol{z}})\boldsymbol{z} + \boldsymbol{C}\boldsymbol{z} + \boldsymbol{\eta}, \quad (6)$$

where $\boldsymbol{z} = [z_1, \dots, z_N]^T$, $\boldsymbol{a} = [a_1, \dots, a_N]^T$, $\boldsymbol{\omega} = [\omega_1, \dots, \omega_N]^T$, $\boldsymbol{\eta} = [\eta_1, \dots, \eta_N]^T$, and $\boldsymbol{S} = [S_1, \dots, S_N]^T$ is a vector containing the connectivity strength of each node; i.e., $S_i = \sum_j C_{ij}$. The superscript $[\ ]^T$ represents the transpose, $\odot$ is the Hadamard element-wise product, and $\bar{\boldsymbol{z}}$ is the complex conjugate of $\boldsymbol{z}$. This equation describes the linear fluctuations around the fixed point $\boldsymbol{z} = 0$, which is the solution of $\frac{d\boldsymbol{z}}{dt} = 0$. Separating the real and imaginary parts of the state variables, and discarding the higher-order terms $(\boldsymbol{z}\odot\bar{\boldsymbol{z}})\boldsymbol{z}$, the evolution of the linear fluctuations follows a Langevin stochastic linear equation:

$$\frac{d}{dt}\delta\boldsymbol{u} = \boldsymbol{J}\delta\boldsymbol{u} + \boldsymbol{\eta}, \quad (7)$$

where the $2N$-dimensional vector $\delta\boldsymbol{u} = [\delta\boldsymbol{x}, \delta\boldsymbol{y}]^{\boldsymbol{T}} = [\delta x_1, \dots, \delta x_N, \delta y_1, \dots, \delta y_N]^T$ contains the fluctuations of real and imaginary state variables. The $2N \times 2N$ matrix $\boldsymbol{J}$ is the Jacobian of the system evaluated at the fixed point, which can be written as a block matrix

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J_{xx}} & \boldsymbol{J_{xy}} \\ \boldsymbol{J_{yx}} & \boldsymbol{J_{yy}} \end{bmatrix}, \quad (8)$$

where $\boldsymbol{J_{xx}}$, $\boldsymbol{J_{xy}}$, $\boldsymbol{J_{yx}}$, $\boldsymbol{J_{yy}}$ are $N \times N$ matrices $\boldsymbol{J_{xx}} = \boldsymbol{J_{yy}} = \mathrm{diag}(\boldsymbol{a} - \boldsymbol{S}) + \boldsymbol{C}$ and $\boldsymbol{J_{xy}} = -\boldsymbol{J_{yx}} = \mathrm{diag}(\boldsymbol{\omega})$, where $\mathrm{diag}(\boldsymbol{v})$ is the diagonal matrix whose diagonal is the vector $\boldsymbol{v}$. We note that the above linearization is only valid if $\boldsymbol{z} = 0$ is a stable solution of the system; i.e., if all eigenvalues of $\boldsymbol{J}$ have negative real part.

C. Quantifying violations of model-based FDT

To examine the violations of FDT, we must first compute the covariance matrix $\boldsymbol{K} = \langle \delta\boldsymbol{u}\delta\boldsymbol{u}^{\boldsymbol{T}} \rangle$. We begin by writing Equation 7 as $d\delta\boldsymbol{u} = \boldsymbol{J}\delta\boldsymbol{u}dt + d\boldsymbol{W}$, where $d\boldsymbol{W}$ is an $2N$-dimensional Wiener process with covariance $\langle d\boldsymbol{W}d\boldsymbol{W}^{\boldsymbol{T}} \rangle = \boldsymbol{Q}dt$ and $\boldsymbol{Q}$ is the noise covariance matrix (which is diagonal if the noise is uncorrelated). Using Itô's stochastic calculus, we get $d(\delta\boldsymbol{u}\delta\boldsymbol{u}^{\boldsymbol{T}}) = d(\delta\boldsymbol{u})\delta\boldsymbol{u}^{\boldsymbol{T}} + \delta\boldsymbol{u}d(\delta\boldsymbol{u}^{\boldsymbol{T}}) + d(\delta\boldsymbol{u})d(\delta\boldsymbol{u}^{\boldsymbol{T}})$. Taking expectations, keeping terms to first order in the differential dt, and noting that $\langle \delta\boldsymbol{u}d\boldsymbol{W}^{\boldsymbol{T}} \rangle = 0$, we obtain:

$$\frac{d\boldsymbol{K}}{dt} = \boldsymbol{J}\,\boldsymbol{K} + \boldsymbol{K}\,\boldsymbol{J}^{\boldsymbol{T}} + \boldsymbol{Q}. \quad (9)$$

Hence, the stationary covariances (for which $\frac{d\boldsymbol{K}}{dt} = 0$) can be obtained by solving the following analytic equation:

$$\boldsymbol{J}\,\boldsymbol{K} + \boldsymbol{K}\,\boldsymbol{J}^{\boldsymbol{T}} + \boldsymbol{Q} = \boldsymbol{0}. \quad (10)$$

This Lyapunov equation can be solved using the eigen-decomposition of the Jacobian matrix $\boldsymbol{J}$ [36]. We then obtained the simulated functional connectivity $\boldsymbol{FC}^{model}$ from the first N rows and columns of the covariance $\boldsymbol{K}$, which corresponds to the real part of the dynamics (precisely representing the BOLD fMRI signal).

Still, even if the analytical solution is possible, in order to fit the model to the empirical data (BOLD fMRI of each participant in each brain state), for the optimization of the coupling connectivity matrix $\boldsymbol{C}$, similar to the work of Gilson and colleagues, here it proved more robust to estimate this numerically by using a pseudo-gradient descent procedure [37, 38]. Specifically, we fit $\boldsymbol{C}$ such that the model optimally reproduces the empirically measured covariances $\boldsymbol{FC}^{empirical}$ (i.e., the normalised covariance matrix of the functional neuroimaging data) and the empirical time-shifted covariances $\boldsymbol{FS}^{empirical}(\tau)$, where τ is the time lag, which are normalized for each pair of regions i and j by $\sqrt{KS_{ii}^{empirical}(0)KS_{jj}^{empirical}(0)}$. We selected the parameter τ, which led to a decrease in the averaged autocorrelation. We note that fitting the time-shifted correlations can lead to asymmetries in the connectivity $\boldsymbol{C}$, which, in turn, can produce non-equilibrium dynamics and violations of the FDT. These normalised time-shifted covariance matrices are generated by taking the shifted covariance matrix $\boldsymbol{KS}^{empirical}(\tau)$ and dividing each pair (i, j) by $\sqrt{KS_{ii}^{empirical}(0)KS_{jj}^{empirical}(0)}$. Note that these normalised time-shifted covariances break the symmetry of the couplings and thus improve the level of fitting [39]. Importantly, the linear approximation allows us an analytical derivation, which means that the estimation of all the functional observables is directly derived without explicit simulation of the time dynamics, that is equivalent to a numerical simulation of infinite duration. We fitted the exact model's parameters to the corresponding empirical observables.

Using a heuristic pseudo-gradient algorithm, we proceeded to update the $\boldsymbol{C}$ until the fit is fully optimised. More specifically, the updating uses the following form:

$$C_{ij} = C_{ij} + \alpha\left(FC_{ij}^{empirical} - FC_{ij}^{model}\right) + \varsigma\left(FS_{ij}^{empirical}(\tau) - FS_{ij}^{model}(\tau)\right), \quad (11)$$

where $FS_{ij}^{model}(\tau)$ is defined similar to $FS_{ij}^{empirical}(\tau)$. In other words it is given by the first N rows and columns of the simulated τ time-shifted covariances $\boldsymbol{KS}^{model}(\tau)$ normalized by dividing each pair (i, j) by $\sqrt{KS_{ii}^{model}(0)KS_{jj}^{model}(0)}$, being $\boldsymbol{KS}^{model}(\tau)$ the shifted simulated covariance matrix computed as following:

$$\boldsymbol{KS}^{model}(\tau) = \exp(\tau\boldsymbol{J})\,\boldsymbol{K}. \quad (12)$$

Note that $\boldsymbol{KS}^{model}(0) = \boldsymbol{K}$. The model was run repeatedly with the updated $\boldsymbol{C}$ until the fit converges towards a stable value. We initialised $\boldsymbol{C}$ using the anatomical connectivity (obtained with probabilistic tractography from dMRI) and only update known existing connections from this matrix (in either hemisphere). However, there is one exception to this rule which is that the algorithm also updates homologue connections between the same regions in either hemisphere, given that tractography is known to be less accurate when accounting for this connectivity. For the Stuart-Landau model, we used $\alpha = \varsigma = 0.00001$ and continue until the algorithm converges. For each iteration we compute the model results as the average over as many simulations as there are participants. Overall, we use the term Generative Effective Connectivity (GEC) for the optimised $\boldsymbol{C}$ [40]. Note, that this generative matrix is asymmetric, since it is able to describe the breaking of the detailed balance in the empirical data. In contrast, the original anatomical SC and functional FC matrices are by definition symmetric and does not provide information on the generative principles. Instead, the FDT framework is capturing the asymmetry of the underlying information flow through the GEC generating brain dynamics in any given brain state.

After fitting an individualized coupling matrix $\boldsymbol{C}$ for each participant and each brain state, we derived an analytical form for the deviation from the FDT (corresponding to Equation 4). First, we derive the expectation values of the state variables $\langle \delta\boldsymbol{u} \rangle_{\varepsilon_j}$ when a perturbation ε is applied to the component j. From Equation 7, we have the relationship $\frac{d}{dt}\langle \delta\boldsymbol{u} \rangle_{\varepsilon_j} = \boldsymbol{J}\langle \delta\boldsymbol{u} \rangle_{\varepsilon_j} + \boldsymbol{h}_j = 0$, where $\boldsymbol{h}_j$ is a $2N$-dimensional vector of all zeros except for the j component,

which is equal to ε. Solving for the desired expectation value, we obtain $\langle \delta \boldsymbol{u} \rangle_{\varepsilon_j} = -\boldsymbol{J}^{-1}\boldsymbol{h}_j$. Defining $\langle \delta \boldsymbol{x} \rangle_j = \langle \delta \boldsymbol{x} \rangle_{\varepsilon_j}/\varepsilon$, i.e. the real part of $\langle \delta \boldsymbol{u} \rangle_j$, we can now derive the deviation from the FDT for region i when a perturbation is applied to region j:

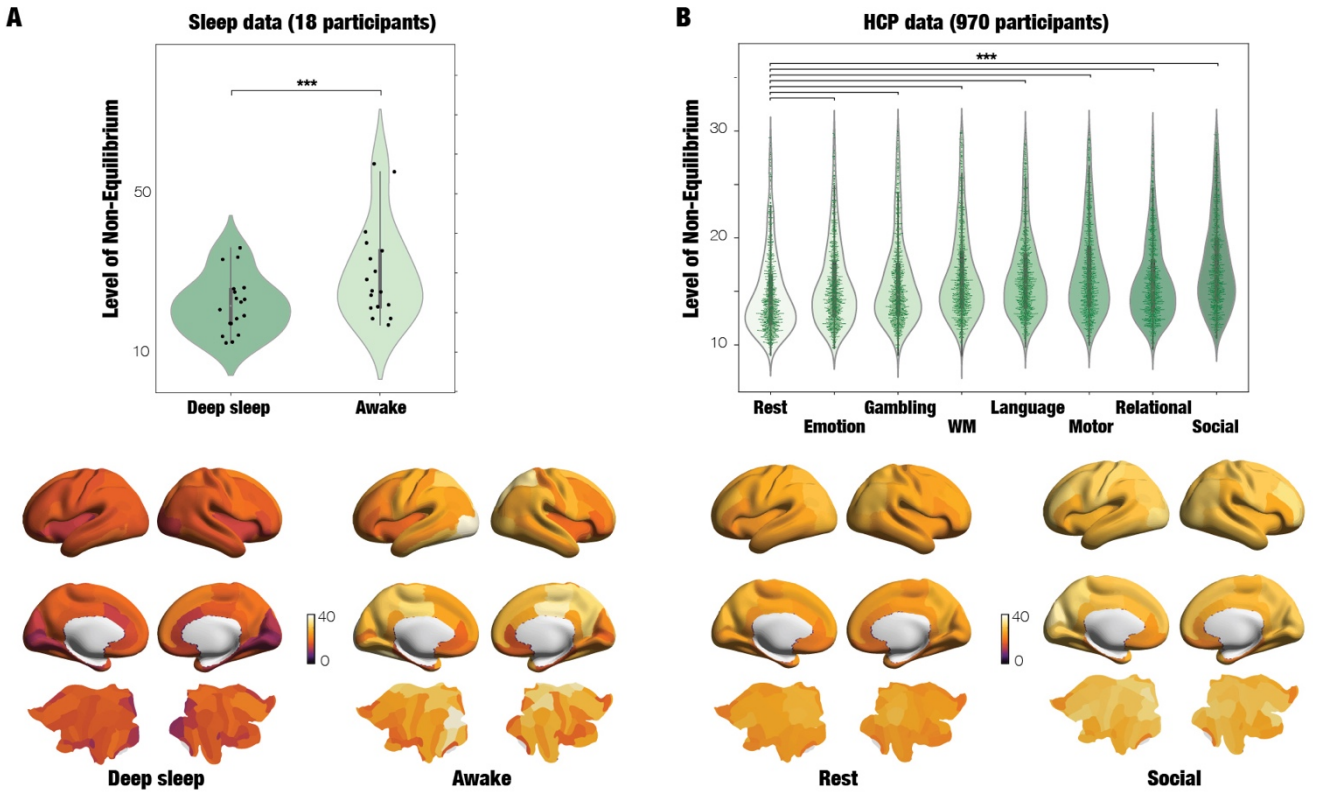


FIG 2. Non-equilibrium fingerprints of brain states captured by deviations from FDT. (a) Significant differences in deviations from FDT were found when comparing deep sleep with wakefulness in neuroimaging data from 18 healthy participants with precise sleep staging using polysomnography ($p<0.001$). Renderings of the resulting perturbability maps on the human brain (3D views of side and midline as well as cortical flatmaps) show more homogeneous and much lower levels of non-equilibrium for deep sleep compared to wakefulness. (b) Similarly, significant differences were observed when comparing resting state with seven different tasks in 970 participants from the human connectome project ($p<0.001$ for all comparisons). As can be seen the perturbability maps are more homogeneous and much lower levels of non-equilibrium for rest compared to task (here for the SOCIAL task)

$$D_{i,j} = \frac{2\langle \delta x_i \delta x \rangle_0/\sigma^2 - \langle \delta x_i \rangle_j}{\langle \delta x_i \rangle_j}, \tag{13}$$

where the term $2/\sigma^2$ plays the role of the inverse temperature β, and the covariance $\langle \delta x_i \delta x \rangle_0$ is derived from $\boldsymbol{KS}^{model}$. For numerical reasons, we quantify the system-wide effect of perturbing the component j by averaging the numerator and denominator over the regions; i.e.,

$$P_j = \frac{\frac{1}{N}\sum_i 2\langle \delta x_i \delta x_j \rangle_0/\sigma^2 - \langle \delta x_i \rangle_j}{\frac{1}{N}\sum_i \langle \delta x_i \rangle_j}. \tag{14}$$

The vector P defines a *perturbability map* over all brain regions in a given brain state. For each participant, the level of non-equilibrium $\widehat{D}$ is finally computed by averaging the deviation from the FDT over all possible perturbations; i.e.,

$$\widehat{D} = \frac{1}{N}\sum_j P_j. \tag{15}$$

IV. RESULTS ON HUMAN EMPIRICAL NEUROIMAGING DATA

We applied this FDT framework to two empirical neuroimaging datasets in humans, with whole-brain activity measured using BOLD fMRI. The first dataset consists of 18 human participants whose sleep stages were precisely characterised by two independent neurologists from simultaneous electroencephalography (EEG) recordings [41]. We considered two stages of consciousness: wakefulness and deep sleep (N3) (see Appendix A for details on the experimental setup and data processing). The second dataset consists of 970 participants from the Human Connectome Project (again, see Appendix A for details), who were recorded during resting state and seven different tasks spanning a broad range of cognitive and emotional processing [42].

A. Wakefulness and deep sleep

First, as shown in Fig 2(a), we applied the FDT framework to the sleep dataset and found significant differences in the deviations from the FDT when comparing deep sleep with wakefulness ($p<0.001$, permutation test). Specifically, we computed $\widehat{D}$ (from Equation 15) for each participant and each level of consciousness, revealing a decrease in violations of the FDT (or level of non-equilibrium) during deep sleep compared with wakefulness. This difference can be interpreted as a flattening of the hierarchical organisation during deep sleep; that is, a brain state with more symmetrical interactions compared with wakefulness. These violations of the FDT can be clearly visualized using the corresponding perturbability maps (vector P_j from Equation 14), which show more homogeneous and much lower levels of non-equilibrium responses for deep sleep compared to wakefulness.

B. Cognitive tasks and resting state dynamics

Second, as shown in Fig 2(b), we observed significant differences in the violations of the FDT when comparing resting state with different cognitive tasks across 970 healthy participants ($p<0.001$ for all comparisons, permutation tests). Just as in the investigations of sleep states, we computed $\widehat{D}$ for each participant and each cognitive task (including rest), revealing differences in the non-equilibrium nature of the brain. For example, the SOCIAL task induced the largest violations of the FDT, reflecting the highest level of non-equilibrium [4, 7, 8]. By contrast, we observed closer agreement with the FDT for resting compared to each of the cognitive tasks. Indeed, the perturbability map for rest is more homogeneous with responses that are closer to equilibrium compared to the SOCIAL task. These results suggest that violations of the FDT (and the distance from equilibrium) increase with computational demands. This can be interpreted in terms of the breaking of the detailed balance, where the flow of information requires asymmetric interactions between brain regions.

C. Hierarchy, asymmetry and violations of FDT

To investigate the mechanisms underlying violations of the FDT, and the relationship to the breaking of detailed balance, we constructed two simple linear models with differing levels of asymmetry in their interactions. Specifically, to relate the asymmetry of the underlying coupling matrix to violations of the FDT, we use a Langevin equation:

$$\frac{d\boldsymbol{b}}{dt} = \boldsymbol{L}\boldsymbol{b} + \boldsymbol{\eta}, \qquad (16)$$

where $\boldsymbol{b} = [b_1, \dots, b_N]^T$ models the bold signal in a parcellation of N regions, $\boldsymbol{L}$ is the coupling matrix, and $\boldsymbol{\eta} = [\eta_1, \dots, \eta_N]^T$ the additive Gaussian noise. We consider two different models, each generated by fitting the couplings $\boldsymbol{L}$ to the empirical neural activity during wakefulness to obtain realistic generative effective connectivity. We first define a symmetric model that is only fit to the equal-time empirical correlations, resulting in symmetric effective connectivity $\boldsymbol{L}$. We then define an asymmetric model that fit to both the equal-time and time-delayed correlations, resulting in asymmetric connectivity.

Figure 3 shows the importance of asymmetric couplings for violations of the FDT. Specifically, Fig 3(a) shows how the linear *symmetric* model generates a fully symmetric connectivity matrix (left panel), which can be observed by computing $|\boldsymbol{L} - \boldsymbol{L}^T|$ (middle panel). Notably, these symmetric couplings yield fully equilibrium dynamics, and therefore do not generate any violations of the FDT (right panel).

By contrast, Fig 3(b) shows the connectivity matrix of the asymmetric model (left panel), which is asymmetric (middle panel) and thus induces significant violations of the FDT (right panel). These differences between the asymmetric and symmetric models are illustrated in Fig 3(c). The first panel shows a scatterplot of the mean regional FDT deviations (i.e., averaging over the rows of the deviation matrix $D_{i,j}$) as a function of the mean regional connectivity strength (with red points for the asymmetrical model and black points for the symmetrical model). In the asymmetric model, we observe a significant correlation (of 0.77) between the FDT deviations and the regional connectivity strength, while this correlation vanishes for the symmetric model. The second panel shows the scatterplot of the mean perturbation site FDT deviation (i.e., averaging over the columns of the deviation matrix $D_{i,j}$) as a function of the mean perturbation site connectivity strength, revealing a negative correlation (-0.87) for the asymmetric model. The third panel shows a violin plot of the significant mean FDT deviation for the symmetric (grey) and asymmetric (green) models across all regions and sites ($p<0.001$, permutation testing).

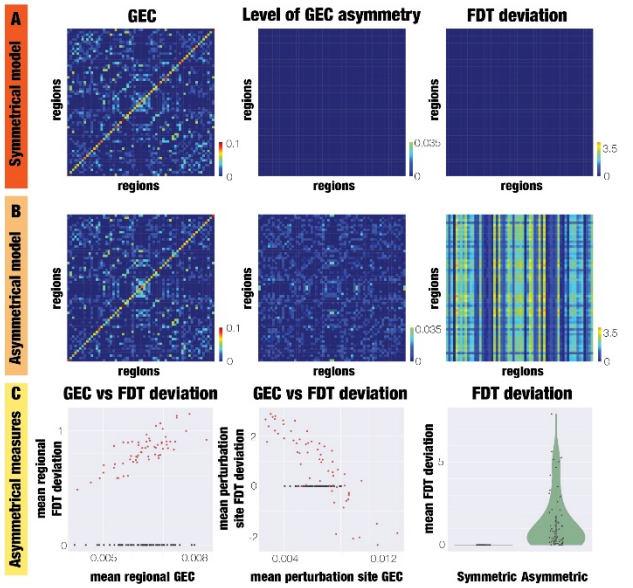


FIG 3. Whole-brain models show the importance of asymmetric coupling compared to symmetric coupling. The breaking of the detailed balance in the brain gives rise to non-equilibrium, linked to the underlying asymmetric coupling, which can be captured with deviations from the FDT. (a) The simplest linear symmetric model generates a Generative Effective Connectivity (GEC) matrix (shown on the left) which is fully symmetrical (shown by the middle matrix) and does not generate any deviations from FDT as shown by the right matrix (level of FDT deviation). (b) In contrast, the asymmetric model generates a GEC matrix (left) which is asymmetric (middle) and with significant deviations from FDT (right matrix). (c) The row shows various measurements of the asymmetry for the asymmetrical (red points) and symmetrical (black points) models. The first panel shows a scatterplot of the mean regional FDT deviation as a function of the mean regional GEC. As can be seen the asymmetrical model show a correlation between FDT deviation and the mean regional GEC, while this is not the case for the symmetrical model. The second panel shows a scatterplot of the mean perturbation site FDT deviation as a function of the mean perturbation site GEC. Again, the asymmetrical model generates a negative correlation between mean perturbation site FDT deviation and the mean perturbation site GEC. This is not the case for the symmetrical model. Finally, the right panel shows a violin plot of the mean FDT deviation for the symmetrical (grey) and asymmetrical (green) models across all regions and sites.

D. Model-based and model-free non-equilibrium metrics

In order to compare our model-based FDT deviation with model-free measurements of the level of non-equilibrium in brain dynamics, we used the INSIDEOUT framework for characterising the arrow of time in brain signals [8]. Using large-scale neuroimaging resting state data from the over 1,000 participants in the Human Connectome Project (HCP), we demonstrated that both measures capture the underlying breaking of detailed balance in the generative space. The INSIDEOUT framework aims to measure the arrow of time and its link with non-equilibrium and time asymmetry. These ideas from statistical physics are applied to brain signals to characterize the level of reversibility/non-equilibrium. Specifically, the measurements are performed directly from

the empirical data without any underlying model assumptions, using time-shifted correlations.

As shown in the scatter plot in Appendix C, we found a strong correlation of 0.75 ($p<0.001$) when comparing all participants using the FDT and INSIDEOUT frameworks. Furthermore, we demonstrated that both measures are highly correlated with the breaking of the detailed balance quantified by the level of asymmetry in generative space (i.e., GEC) computed as the mean of $|\boldsymbol{C} - \boldsymbol{C^T}|$. We obtained a correlation of 0.91 ($p<0.001$) between FDT and asymmetry of GEC, and a correlation of 0.82 ($p<0.83$) between INSIDEOUT and asymmetry of GEC. In summary, we demonstrate that both the model-based FDT and the model-free INSIDEOUT framework capture the level of non-equilibrium underlying empirical neuroimaging data.

V. CONCLUSION

For the first time, we applied the FDT to neural activity fitted by a whole-brain model, which allowed us to investigate how non-equilibrium dynamics are associated with sleep, wakefulness and seven cognitive tasks. We find that violations of the FDT (and thus divergences from equilibrium) are driven by asymmetries in the couplings between brain regions, thus revealing the role of hierarchical organization in non-equilibrium dynamics. The largest violations of the FDT were observed when subjects performed cognitive tasks (with the SOCIAL task inducing the largest violations), while the neural dynamics were closer to equilibrium for sleep than wakefulness. These differences directly reflect the computational demands that require asymmetric information flow between brain regions, thus breaking detailed balance and promoting non-equilibrium dynamics. Schrödinger hypothesized that this increasing asymmetrical information flow is important for sustaining life [43], and here we extend this thermodynamic principle to neural computations.

Using thermodynamics to describe brain dynamics is an emerging field [19], which has already yielded important new insights into the non-equilibrium nature of brain function [1, 4, 7, 8]. Excitingly, these insights include the demonstration of how the arrow of time, or irreversibility, of brain signals can shed new light on the definition of brain states [7, 8, 40]. Interestingly, the FDT was recently derived in a spiking neuron model to assess the statistics of the unknown fluctuations of the neuronal dynamics [44] and for stochastic oscillators [45]. Meanwhile, brain dynamics have also been shown to be turbulent [46, 47], allowing the fast information transfer needed for time-critical decisions in the brain (MEG).

We note that non-equilibrium dynamics are found at different spatial and temporal scales [48]. For instance, while larger and more complex biological structures, such as cells, might appear to be at equilibrium, they are sustained by non-equilibrium processes at small scales (e.g., molecules) [49]. In particular, overwhelming evidence suggests that neural dynamics at microscopical and fast time scales are intrinsically non-equilibrium phenomena based on the fact that cellular and molecular functions consume energy [50-53]. Despite the clear importance of non-equilibrium dynamics at this microscale, the role of broken detailed balance in the brain as a macroscopic system composed of many interacting components has been less investigated. Our approach directly demonstrates the role of broken detailed balance in the large-scale whole-brain activity by quantifying the violation of the FDT, which is independently relevant to the non-equilibrium dynamics observed at the microscopic and short time scale.

Overall, the model-based FDT approach introduced here holds great promise for revealing the underlying principles of non-equilibrium dynamics in the human brain. Specifically, this approach provides a necessary framework for the very influential papers by Massimini and colleagues. They ran a series of groundbreaking experiments using transcranial magnetic stimulation (TMS) and electroencephalography (EEG) to measure fluctuations and dissipation after perturbations [2, 5, 6]. The resulting Perturbational Complexity Index (PCI) measures the amount of information contained in the amplitude of the average perturbation-evoked responses by calculating the Lempel-Ziv complexity of the binary matrix describing the statistically significant sources, in space and time, of the EEG signals [2]. This PCI measure has been successfully used for separation of brain states in healthy subjects during wakefulness, dreaming, sleep, under different levels of anaesthesia and in coma [2, 5, 6]. The present work provides a theoretical framework not only to explain these important findings in terms of non-equilibrium dynamics but also provides principled novel measurements for predicting system-wide response of the brain to any targeted perturbations in non-equilibrium brain state, whether in health or disease. The main advantage of the FDT framework over previous empirical approaches is that it can provide deep insights into the causal generative mechanisms of brain dynamics in any brain state, and avoids expensive experiments, improves statistical robustness and minimises potential ethical concerns. As such through the use of the FDT framework on the abundant fMRI data from different brain states, we may inch closer to a useful definition of brain states and potential insights into how to transition between them, and thus, for example, providing new useful information on how best to wake comatose patients.

ACKNOWLEDGEMENTS

G.D. was supported by a number of sources: NEMESIS project (ref. 101071900) funded by the EU ERC Synergy Horizon Europe; AGAUR research support grant (ref. 2021 SGR 00917) funded by the Department of Research and Universities of the Generalitat of Catalunya; project PID2022-136216NB-I00 financed by the MCIN /AEI /10.13039/501100011033 / FEDER, UE., the Ministry of Science and Innovation, the State Research Agency and the European Regional Development Fund. C.W.L. is supported by a James S. McDonnell Foundation Postdoctoral Fellowship Award and by the National Science Foundation through the Center for the Physics of Biological Function (PHY–1734030). Y.S.P is supported by European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant 896354. M.L.K. is supported by the Center for Music in the Brain, funded by the Danish National Research Foundation (DNRF117), and Centre for Eudaimonia and Human Flourishing at Linacre College funded by the Pettit and Carlsberg Foundations.

[1] Y. Sanz Perl, H. Bocaccio, I. Perez-Ipina, S. Laureys, H. Laufs, M. L. Kringelbach, G. Deco, and E. Tagliazucchi, *Non-equilibrium brain dynamics as a signature of consciousness*, Physical Review E **104** 014411 (2021).

[2] A. G. Casali *et al.*, *A theoretically based index of consciousness independent of sensory processing and behavior*, Science translational medicine **5** 198ra105 (2013).

[3] A. Seif, M. Hafezi, and C. Jarzynski, *Machine learning the thermodynamic arrow of time*, Nat. Physics **17** 105 (2021).

[4] C. W. Lynn, E. J. Cornblath, L. Papadopoulos, M. A. Bertolero, and D. S. Bassett, *Broken detailed balance and entropy production in the human brain*, Proc. Natl. Acad. Sci. U.S.A. **118** (2021).

[5] F. Ferrarelli, M. Massimini, S. Sarasso, A. Casali, B. A. Riedner, G. Angelini, G. Tononi, and R. A. Pearce, *Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness*, Proc. Natl. Acad. Sci. U.S.A. **107** 2681 (2010).

[6] M. Massimini, F. Ferrarelli, R. Huber, S. K. Esser, H. Singh, and G. Tononi, *Breakdown of cortical effective connectivity during sleep*, Science **309** 2228 (2005).

[7] G. Deco, Y. Sanz Perl, L. de la Fuente, J. Sitt, B. T. T. Yeo, E. Tagliazucchi, and M. L. Kringelbach, *The arrow of time of brain signals in cognition: Potential intriguing role of parts of the default mode network*, Network neuroscience **7** 966 (2023).

[8] G. Deco, Y. Sanz Perl, E. Tagliazucchi, and M. L. Kringelbach, *The INSIDEOUT framework provides precise signatures of the balance of intrinsic and extrinsic dynamics in brain states*, Commun. Biol. **5** 572 (2022).

[9] A. Crisanti and F. Ritort, *Violation of the fluctuation–dissipation theorem in glassy systems: basic notions and the numerical evidence*, Journal of Physics A: Mathematical and General **36** R181 (2003).

[10] L. Onsager, *Reciprocal Relations in Irreversible Processes. II*, Physical Review **38** 2265 (1931).

[11] L. Onsager, *Reciprocal Relations in Irreversible Processes. I*, Physical Review **37** 405 (1931).

[12] M. L. Kringelbach and G. Deco, *Brain States and Transitions: Insights from Computational Neuroscience*, Cell Reports **32** 108128 (2020).

[13] G. Deco and M. L. Kringelbach, *Great Expectations: Using Whole-Brain Computational Connectomics for Understanding Neuropsychiatric Disorders*, Neuron **84** 892 (2014).

[14] G. Deco, G. Tononi, M. Boly, and M. L. Kringelbach, *Rethinking segregation and integration: contributions of whole-brain modelling* Nat Rev Neurosci **16** 430 (2015).

[15] M. Breakspear, *Dynamic models of large-scale brain activity*, Nat. Neurosci. **20** 340 (2017).

[16] M. Breakspear, *"Dynamic" connectivity in neural systems: theoretical and empirical considerations*, Neuroinformatics **2** 205 (2004).

[17] A. Ghosh, Y. Rho, A. R. McIntosh, R. Kotter, and V. K. Jirsa, *Noise during rest enables the exploration of the brain's dynamic repertoire*, PLoS Comput. Biol. **4** e1000196 (2008).

[18] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns, *Network structure of cerebral cortex shapes functional connectivity on multiple time scales*, Proc. Natl. Acad. Sci. U.S.A. **104** 10240 (2007).

[19] C. W. Lynn and D. S. Bassett, *The physics of brain network structure, function and control*, Nature Reviews Physics **1** 318 (2019).

[20] G. Deco, V. K. Jirsa, and A. R. McIntosh, *Emerging concepts for the dynamical organization of resting state activity in the brain*, Nat. Rev. Neurosci. **12** 43 (2011).

[21] G. Deco, A. Ponce-Alvarez, D. Mantini, G. L. Romani, P. Hagmann, and M. Corbetta, *Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations*, J. Neurosci. **33** 11239 (2013).

[22] Y. A. Kuznetsov, *Elements of applied bifurcation theory* (Springer, New York, 1998).

[23] F. Freyer, J. A. Roberts, R. Becker, P. A. Robinson, P. Ritter, and M. Breakspear, *Biophysical mechanisms of multistability in resting-state cortical rhythms*, J. Neurosci. **31** 6353 (2011).

[24] F. Freyer, J. A. Roberts, P. Ritter, and M. Breakspear, *A canonical model of multistability and scale-invariance in biological systems*, PLoS Comput. Biol. **8** e1002634 (2012).

[25] G. Deco, J. Cabral, M. Woolrich, A. B. A. Stevner, T. Van Hartevelt, and M. L. Kringelbach, *Single or Multi-Frequency Generators in On-going MEG Data: a Mechanistic Whole-Brain Model of empirical MEG data*, Neuroimage **152** 538 (2017).

[26] G. Deco, J. Cruzat, J. Cabral, E. Tagliazucchi, H. Laufs, N. K. Logothetis, and M. L. Kringelbach, *Awakening: predicting external stimulation forcing transitions between different brain states*, PNAS **116** 18088 (2019).

[27] M. L. Kringelbach, J. Cruzat, J. Cabral, G. M. Knudsen, R. L. Carhart-Harris, P. C. Whybrow, N. K. Logothetis, and G. Deco, *Dynamic Coupling of Whole-Brain Neuronal and Neurotransmitter Systems*, PNAS in press (2020).

[28] G. Deco, M. L. Kringelbach, V. Jirsa, and P. Ritter, *The Dynamics of Resting Fluctuations in the Brain: Metastability and its Dynamical Core [bioRxiv 065284]*, Scientific Reports **7** 3095 (2017).

[29] D. R. Chialvo, *Emergent complex neural dynamics*, Nat. Physics **6** 744 (2010).

[30] A. Haimovici, E. Tagliazucchi, P. Balenzuela, and D. R. Chialvo, *Brain organization into resting state networks emerges at criticality on a model of the human connectome*, Physical review letters **110** 178101 (2013).

[31] G. Deco and V. K. Jirsa, *Ongoing cortical activity at rest: criticality, multistability, and ghost attractors*, J. Neurosci. **32** 3366 (2012).

[32] Y. Sanz Perl, A. Escrichs, E. Tagliazucchi, M. L. Kringelbach, and G. Deco, *Strength-dependent perturbation of whole-brain model working in different regimes reveals the role of fluctuations in brain dynamics*, PLoS Comput Biol **18** e1010662 (2022).

[33] H. Bocaccio, C. Pallavicini, M. N. Castro, S. M. Sanchez, G. De Pino, H. Laufs, M. F. Villarreal, and E. Tagliazucchi, *The avalanche-like behaviour of large-scale haemodynamic activity from wakefulness to deep sleep*, J. R. Soc. Interface **16** 20190262 (2019).

[34] I. P. Ipina, P. D. Kehoe, M. Kringelbach, H. Laufs, A. Ibanez, G. Deco, Y. S. Perl, and E. Tagliazucchi, *Modeling regional changes in dynamic stability during sleep and wakefulness*, Neuroimage **215** 116833 (2020).

[35] B. Jobst, H. Hindriks, H. Laufs, E. Tagliazucchi, G. Hahn, A. Ponce-Alvarez, A. B. A. Stevner, M. L. Kringelbach, and G. Deco, *Increased stability and breakdown of brain effective connectivity during slow-*

wave sleep: mechanistic insights from whole-brain computational modelling, Scientific Reports in press (2017).

[36] G. Deco, A. Ponce-Alvarez, P. Hagmann, G. L. Romani, D. Mantini, and M. Corbetta, *How local excitation-inhibition ratio impacts the whole brain dynamics*, J. Neurosci. **34** 7886 (2014).

[37] M. Gilson *et al.*, *Model-based whole-brain effective connectivity to study distributed cognition in health and disease*, Netw. Neurosci. **4** 338 (2020).

[38] M. Gilson, R. Moreno-Bote, A. Ponce-Alvarez, P. Ritter, and G. Deco, *Estimation of Directed Effective Connectivity from fMRI Functional Connectivity Hints at Asymmetries of Cortical Connectome*, PLoS Comput. Biol. **12** e1004762 (2016).

[39] M. Gilson, G. Deco, K. J. Friston, P. Hagmann, D. Mantini, V. Betti, G. L. Romani, and M. Corbetta, *Effective connectivity inferred from fMRI transition dynamics during movie viewing points to a balanced reconfiguration of cortical interactions*, NeuroImage **180** 534 (2017).

[40] M. L. Kringelbach, Y. Sanz Perl, E. Tagliazucchi, and G. Deco, *Toward naturalistic neuroscience: Mechanisms underlying the flattening of brain hierarchy in movie-watching compared to rest and task*, Science Advances in press (2023).

[41] E. Tagliazucchi and H. Laufs, *Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep*, Neuron **82** 695 (2014).

[42] D. M. Barch *et al.*, *Function in the human connectome: task-fMRI and individual differences in behavior*, NeuroImage **80** 169 (2013).

[43] E. Schrödinger, *What is life? The Physical Aspect of the Living Cell* (Cambridge University Press, Cambridge, 1944).

[44] B. Lindner, *Fluctuation-Dissipation Relations for Spiking Neurons*, PRL **129** 198101 (2022).

[45] A. Pérez-Cervera, B. Gutkin, P. J. Thomas, and N. Lindner, *A Universal Description of Stochastic Oscillators*, arXiv 2303.03198 (2023).

[46] G. Deco and M. L. Kringelbach, *Turbulent-like dynamics in the human brain*, Cell Reports **33** 108471 (2020).

[47] G. Deco, Y. Sanz Perl, P. Vuust, E. Tagliazucchi, H. Kennedy, and M. L. Kringelbach, *Rare long-range cortical connections enhance human information processing*, Curr. Biol. **31** 1 (2021).

[48] M. Esposito, *Stochastic thermodynamics under coarse graining*, Physical review. E, Statistical, nonlinear, and soft matter physics **85** 041125 (2012).

[49] D. A. Egolf, *Equilibrium regained: from nonequilibrium chaos to statistical mechanics*, Science **287** 101 (2000).

[50] P. Fries, D. Nikolić, and W. Singer, *The gamma cycle*, Trends Neurosci. **30** 309 (2007).

[51] C. P. Brangwynne, G. H. Koenderink, F. C. MacKintosh, and D. A. Weitz, *Cytoplasmic diffusion: molecular motors mix it up*, J Cell Biol **183** 583 (2008).

[52] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu, *The energy-speed-accuracy tradeoff in sensory adaptation*, Nat Phys **8** 422 (2012).

[53] H. Yin, I. Artsimovitch, R. Landick, and J. Gelles, *Nonequilibrium mechanism of transcription termination from observations of single RNA polymerase molecules*, Proc. Natl. Acad. Sci. U.S.A. **96** 13124 (1999).

[54] R. S. Desikan *et al.*, *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest*, NeuroImage **31** 968 (2006).

[55] K. Setsompop *et al.*, *Pushing the limits of in vivo diffusion MRI for the Human Connectome Project*, NeuroImage **80** 220 (2013).

[56] A. Horn, W. J. Neumann, K. Degen, G. H. Schneider, and A. A. Kuhn, *Toward an electrophysiological "sweet spot" for deep brain stimulation in the subthalamic nucleus*, Hum. Brain Mapp. **38** 3377 (2017).

[57] A. Horn and F. Blankenburg, *Toward a standardized structural-functional group connectome in MNI space*, NeuroImage **124** 310 (2016).

[58] K. H. Maier-Hein *et al.*, *The challenge of mapping the human connectome based on diffusion tractography*, Nat. Commun. **8** 1349 (2017).

[59] K. G. Schilling, A. Daducci, K. Maier-Hein, C. Poupon, J. C. Houde, V. Nath, A. W. Anderson, B. A. Landman, and M. Descoteaux, *Challenges in diffusion MRI tractography - Lessons learned from international benchmark competitions*, Magn Reson Imaging **57** 194 (2019).

[60] M. F. Glasser *et al.*, *The minimal preprocessing pipelines for the Human Connectome Project*, NeuroImage **80** 105 (2013).

[61] S. M. Smith *et al.*, *Resting-state fMRI in the Human Connectome Project*, NeuroImage **80** 144 (2013).

[62] T. Navarro Schroder, K. V. Haak, N. I. Zaragoza Jimenez, C. F. Beckmann, and C. F. Doeller, *Functional topography of the human entorhinal cortex*, eLife **4** e06738 (2015).

[63] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith, *Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers*, NeuroImage **90** 449 (2014).

[64] L. Griffanti *et al.*, *ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging*, NeuroImage **95** 232 (2014).

[65] R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen, *FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data*, Comput. Intell. Neurosci. **2011** 156869 (2011).

[66] A. Klein and J. Tourville, *101 labeled brain images and a consistent human cortical labeling protocol*, Front. Neurosci. **6** 171 (2012).

[67] A. B. A. Stevner *et al.*, *Discovery of key whole-brain transitions and dynamics during human wakefulness and non-REM sleep*, Nat. Commun. **10** 1035 (2019).

[68] G. H. Glover, T. Q. Li, and D. Ress, *Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR*, Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine **44** 162 (2000).

[69] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, *Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain*, NeuroImage **15** 273 (2002).

[70] P. J. Allen, G. Polizzi, K. Krakow, D. R. Fish, and L. Lemieux, *Identification of EEG events in the MR scanner: the problem of pulse artifact and a method for its subtraction*, NeuroImage **8** 229 (1998).

APPENDIX A
PARCELLATION

Both datasets used timeseries from the Mindboggle-modified Desikan-Killiany parcellation [54] with a total of 62 cortical regions (31 regions per hemisphere).

HUMAN CONNECTOME PROJECT: ACQUISITION AND PRE-PROCESSING

Ethics

The Washington University–University of Minnesota (WU-Minn HCP) Consortium obtained full informed consent from all participants, and research procedures and ethical guidelines were followed in accordance with Washington University institutional review board approval (Mapping the Human Connectome: Structure, Function, and Heritability; IRB # 201204036).

Participants

The data set used for this investigation was selected from the March 2017 public data release from the Human Connectome Project (HCP) where we chose a sample of 1003 participants, all of whom have resting state data. For the seven tasks, HCP provides the following numbers of participants: WM =999; SOCIAL=996; MOTOR=996; LANGUAGE=997; GAMBLING=1000; EMOTION=992; RELATIONAL=989. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications using the full HCP dataset.

The HCP task battery of seven tasks

The HCP task battery consists of seven tasks: working memory, motor, gambling, language, social, emotional, relational, which are described in details on the HCP website [42]. HCP states that the tasks were designed to cover a broad range of human cognitive abilities in seven major domains that sample the diversity of neural systems 1) visual, motion, somatosensory, and motor systems, 2) working memory, decision-making and cognitive control systems; 3) category-specific representations; 4) language processing; 5) relational processing; 6) social cognition; and 7) emotion processing. In addition to resting state scans, all 1003 HCP participants performed all tasks in two separate sessions (first session: working memory, gambling and motor; second session: language, social cognition, relational processing and emotion processing).

3T structural data

The HCP structural data were acquired using a customized 3 Tesla Siemens Connectom Skyra scanner with a standard Siemens 32-channel RF-receive head coil. For each participant, at least one 3D T1w MPRAGE image and one 3D T2w SPACE image were collected at 0.7 mm isotropic resolution.

3T diffusion MRI

In order to reconstruct a high-quality structural connectivity (SC) matrix for constructing the whole-brain model (using the DK62 parcellation), we obtained multi-shell diffusion-weighted imaging data from 32 participants from the HCP database (scanned for approximately 89 minutes). The acquisition parameters are described in detail on the HCP website [55]. We estimated the connectivity using the method described by Horn and colleagues [56]. Briefly, the data was processed using a generalized q-sampling imaging algorithm implemented in DSI studio (http://dsi-studio.labsolver.org). Segmentation of the T2-weighted anatomical images produced a white-matter mask and co-registering the images to the b0 image of the diffusion data using SPM12. In each HCP participant, 200,000 fibres were sampled within the white-matter mask. Fibres were transformed into MNI space using Lead-DBS [57]. The methods used the algorithms for false-positive fibres shown to be optimal in recent open challenges [58, 59]. The risk of false positive tractography was reduced in several ways. Most importantly, this used the tracking method achieving the highest (92%) valid connection score among 96 methods submitted from 20 different research groups in a recent open competition [58].

Neuroimaging acquisition for fMRI HCP

The 1003 HCP participants were scanned on a 3-T fMRI using a customized 3-T Siemens Connectom Skyra scanner with a standard Siemens 32-channel RF receive head coil, with the following parameters: 2.0-mm isotropic voxels, TR = 720 ms (time resolution of fMRI data), echo time (TE) = 33.1 ms, flip angle = 52°, field of view (FOV) = 208 × 180 mm, 72 slices, and multiband factor = 8. We used one resting state fMRI acquisition of approximately 15 minutes acquired on the same day, with eyes open with relaxed fixation on a projected bright cross-hair on a dark background as well as data from the seven tasks. The HCP website (http://www.humanconnectome.org/) provides the full details of participants, the acquisition protocol and pre-processing of the data for both resting state and the seven tasks. The time duration of fMRI recordings was: 1200 volumes (resting state); 176 volumes (emotion); 253 volumes (gambling); 405 volumes (WM); 316 volumes (language); 232 volumes (relational); 284 volumes (Motor); and 274 volumes (social). Below we have briefly summarised these.

The pre-processing of the HCP resting state and task datasets is described in details on the HCP website. Briefly, the data is pre-processed using the HCP pipeline which is using standardized methods using FSL (FMRIB Software Library), FreeSurfer, and the Connectome Workbench software [60, 61]. This standard pre-processing included correction for spatial and gradient distortions and head motion, intensity normalization and bias field removal, registration to the T1 weighted structural image, transformation to the 2mm Montreal Neurological Institute (MNI) space, and using the FIX artefact removal procedure [61, 62]. The head motion parameters were regressed out and structured artefacts were removed by ICA+FIX processing (Independent Component Analysis followed by FMRIB's ICA-based X-noiseifier [63, 64]). Pre-processed timeseries of all grayordinates are in HCP CIFTI grayordinates standard space and available in the surface-based CIFTI file for

each participants for resting state and each of the seven tasks.

We used a custom-made Matlab script using the ft_read_cifti function (Fieldtrip toolbox [65]) to extract the average timeseries of all the grayordinates in each region of the Mindboggle-modified Desikan-Killiany parcellation [54] with a total of 62 cortical regions (31 regions per hemisphere) [66], which are defined in the HCP CIFTI grayordinates standard space. The BOLD timeseries were filtered using a second-order Butterworth filter in the range of 0.008-0.08Hz.

HUMAN SLEEP DATA: ACQUISITION AND PRE-PROCESSING

Ethics

Written informed consent was obtained, and the study was approved by the ethics committee of the Faculty of Medicine at the Goethe University of Frankfurt, Germany.

Participants

We used fMRI- and PSG data from 18 participants taken from a larger database that reached all four stages of PSG [41, 67]. Exclusion criteria focussed on the quality of the concomitant acquisition of EEG, EMG, fMRI, and physiological recordings.

Acquisition and pre-processing of fMRI and polysomnography data

Neuroimaging fMRI was acquired on a 3 T system (Siemens Trio, Erlangen, Germany) with the following settings: 1505 volumes of T2*-weighted echo planar images with a repetition time (TR) of 2.08 seconds (time resolution of fMRI data), and an echo time of 30 ms; matrix 64 x 64, voxel size 3 x 3 x 2 mm^3, distance factor 50%, FOV 192 mm^2.

The EPI data were realigned, normalised to MNI space, and spatially smoothed using a Gaussian kernel of 8 mm^3 FWHM in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). Spatial downsampling was then performed to a 4 x 4 x 4 mm resolution. From the simultaneously recorded ECG and respiration, cardiac- and respiratory-induced noise components were estimated using the RETROICOR method [68], and together with motion parameters these were regressed out of the signals. The data were temporally band-pass filtered in the range 0.008-0.08 Hz using a sixth-order Butterworth filter. We extracted the timeseries in the DK62 parcellation [69].

Simultaneous PSG was performed through the recording of EEG, EMG, ECG, EOG, pulse oximetry, and respiration. EEG was recorded using a cap (modified BrainCapMR, Easycap, Herrsching, Germany) with 30 channels, of which the FCz electrode was used as reference. The sampling rate of the EEG was 5 kHz, and a low-pass filter was applied at 250 Hz. MRI and pulse artefact correction were applied based on the average artefact subtraction method[70] in Vision Analyzer2 (Brain Products, Germany). EMG was collected with chin and tibial derivations, and as the ECG and EOG recorded bipolarly at a sampling rate of 5 kHz with a low-pass filter at 1 kHz. Pulse oximetry was collected using the Trio scanner, and respiration with MR-compatible devices (BrainAmp MR+, BrainAmp ExG; Brain Products, Gilching, Germany).

Participants were instructed to lie still in the scanner with their eyes closed and relax. Sleep classification was performed by a sleep expert based on the EEG recordings in accordance with the AASM criteria (2007). Results using the same data and the same pre-processing has previously been reported [41, 67]. We used all the available data for the analysis. The time duration of fMRI signal in each condition is dependent on the time that each participant spent in each sleep stage. Here, we considered the wakefulness and deep sleep stages, where the time duration in each condition varies from 110 to 720 volumes and from 88 to 1002 volumes, respectively.

STATISTICAL COMPARISONS

Differences in probabilities of occurrence before and after injection were statistically assessed using a permutation-based paired t-test. This non-parametric test uses permutations of group labels to estimate the null distribution, which is computed independently for each experimental condition. For each of 1,000 permutations, a t-test is applied to compare populations and a p-value is returned.

APPENDIX B

DERIVATION OF THE FLUCTUATION-DISSIPATION THEOREM IN SPIN SYSTEMS

The simplest thermodynamic model of a system with multiple interacting components is the Ising model. In the Ising model, each component (or spin) is represented by a binary variable x_i. The probability of finding the entire system in state $x = \{x_i\}$ is given by the Boltzmann distribution

$$P(x) = \frac{1}{Z}\exp\left[\beta\left(\sum_{i,j} J_{ij}x_i x_j + \sum_i h_i x_i\right)\right], \quad \text{(B1)}$$

Where β is the inverse temperature, $J_{ij} = J_{ji}$ represents the strength of the interaction between components i and j, h_i is the external influence on component i, and

$$Z = \sum_x \exp\left[\beta\left(\sum_{i,j} J_{ij}x_i x_j + \sum_i h_i x_i\right)\right], \quad \text{(B2)}$$

is the normalization constant (often referred to as the partition function). Because the interactions J_{ij} are symmetric, the system is in equilibrium and the fluctuation-dissipation theorem should hold.

To derive the fluctuation-dissipation theorem, we would like to know how the average state of component i

$$\langle x_i \rangle = \sum_x x_i P(x), \quad \text{(B3)}$$

changes due to a small perturbation h_j coupled to component j. In particular, we have

$$\chi_{ij} = \frac{\partial\langle x_i\rangle}{\partial h_j} = \sum_x x_i \frac{\partial}{\partial h_j} P(x) = \frac{1}{Z}\sum_x x_i \frac{\partial}{\partial h_j}\exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right] - \frac{1}{Z^2}\frac{\partial Z}{\partial h_j}\sum_x x_i \exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right], \quad \text{(B4)}$$

For the first term, we have

$$\frac{1}{Z}\sum_x x_i \frac{\partial}{\partial h_j}\exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right] = \frac{\beta}{Z}\sum_x x_i x_j \exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right] = \beta\sum_x x_i x_j P(x) = \beta\langle x_i x_j\rangle, \quad \text{(B5)}$$

For the second term, we first note that

$$\frac{1}{Z}\frac{\partial Z}{\partial h_j} = \frac{1}{Z}\sum_x \frac{\partial}{\partial h_j}\exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right] = \frac{\beta}{Z}\sum_x x_j \exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right] = \beta\sum_x x_j P(x) = \beta\langle x_j\rangle, \quad \text{(B6)}$$

and so

$$\frac{1}{Z^2}\frac{\partial Z}{\partial h_j}\sum_x x_i \exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right] = \frac{\beta\langle x_j\rangle}{Z}\sum_x x_i \exp\left[\beta\left(\sum_{k,l} J_{kl}x_k x_l + \sum_k h_k x_k\right)\right] = \beta\langle x_i\rangle\langle x_j\rangle \quad \text{(B7)}$$

Thus, putting terms together, we have

$$\chi_{ij} = \frac{\partial\langle x_i\rangle}{\partial h_j} = \beta\left(\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle\right), \quad \text{(B8)}$$

We therefore find that the average response of component i to a perturbation on component j is equal to the spontaneous equilibrium correlation between i and j (scaled by the inverse temperature β). This is precisely the fluctuation-dissipation theorem for the equilibrium Ising model.

APPENDIX C

CORRELATION BETWEEN FDT AND INSIDEOUT

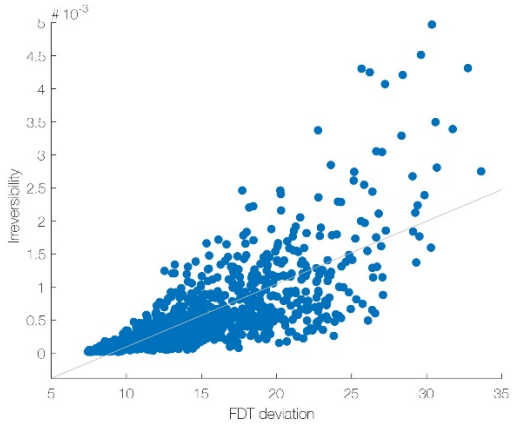


FIG 4. The scatter plot shows a correlation of 0.75 ($p<0.001$) when comparing all HCP participants using the FDT deviation and INSIDEOUT irreversibility measures.