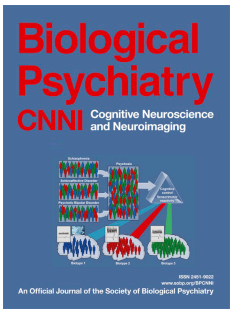
Journal Pre-proof




# Psychiatric Illnesses as Disorders of Network Dynamics

Daniel Durstewitz, Quentin J.M. Huys, Georgia Koppe

PII: S2451-9022(20)30019-7

DOI: https://doi.org/10.1016/j.bpsc.2020.01.001

Reference: BPSC 543

To appear in: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*

Received Date: 17 December 2019

Accepted Date: 6 January 2020

Please cite this article as: Durstewitz D., Huys Q.J.M. & Koppe G., Psychiatric Illnesses as Disorders of Network Dynamics, *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2020), doi: https://doi.org/10.1016/j.bpsc.2020.01.001.

# Psychiatric Illnesses as Disorders of Network Dynamics

Running head: Mental Illnesses as Disorders of Network Dynamics

Daniel Durstewitz[1,2*], Quentin J.M. Huys[3,4], Georgia Koppe[1,5*]

[1]Department of Theoretical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Germany

[2]Faculty of Physics and Astronomy, Heidelberg University, Germany

[3]Division of Psychiatry and Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

[4]Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich

[5]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Germany

[*]corresponding authors:

Daniel Durstewitz

Department of Theoretical Neuroscience

Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University,

J5, 68158 Mannheim, Tel. +49-621-1703-2361, daniel.durstewitz@zi-mannheim.de,

Georgia Koppe

Department of Theoretical Neuroscience & Department of Psychiatry and Psychotherapy

Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University,

J5, 68158 Mannheim, Tel. +49-621-1703-2361, georgia.koppe@zi-mannheim.de

Keywords:
dynamical systems, attractor, chaos, recurrent neural networks, machine learning, schizophrenia

**Abstract**

This review provides a dynamical systems perspective on mental illness. After a brief introduction to the theory of dynamical systems, we focus on the common assumption in theoretical and computational neuroscience that phenomena at subcellular, cellular, network, cognitive and even societal levels could be described and explained in terms of dynamical systems theory (DST). As such, DST may also provide a framework for understanding mental illnesses. The review examines a number of core dynamical systems phenomena and relates each of these to aspects of mental illnesses. This provides an outline of how a broad set of phenomena in serious and common mental illnesses and neurological conditions can be understood in dynamical systems terms. It suggests that the dynamical systems level may provide a central, hub-like level of convergence which unifies and links multiple biophysical and behavioral phenomena, in the sense that diverse biophysical changes can give rise to the same dynamical phenomena and, vice versa, similar changes in dynamics may yield different behavioral symptoms depending on the brain area where these changes manifest. We will also briefly outline current methodological approaches for inferring dynamical systems from data such as EEG, fMRI or self-reports, and discuss the implications of a dynamical view for the diagnosis, prognosis, and treatment of psychiatric conditions. We argue that a consideration of dynamics could play a potentially transformative role in the choice and target of interventions.

# 1. Introduction

Mental illnesses are highly complex, temporally dynamic phenomena (1). Variables across a vast range of timescales – from milliseconds to generations – and levels – from subcellular to societal – interact in complex manners to result in the dynamic, rich and extraordinarily heterogeneous temporal trajectories that are characteristic of the personal and psychiatric histories evident in mental health services across the world. The dynamic and complex nature of these phenomena represents a substantial challenge to our ability to understand mental illnesses, and to treat them. The neglect of the temporal aspects of these phenomena may in part be due to the fact that longitudinal studies have traditionally been more challenging, and hence much research has focused on cross-sectional samples. However, variation observed between individuals will only rarely be informative about individual variation over time (2), and it is arguably the latter that matters more in treatment settings. Time, we suggest, matters, and these dynamical aspects can and need to be addressed directly.

When multiple variables interact with each other in a complex manner over time, then this gives rise to dynamical systems that obey certain rules regardless of the particular nature of the variables involved. The behavior of such systems is studied in the mathematical framework of Dynamical Systems Theory (DST). DST formalizes the complex interaction of variables by a set of differential (if formulated in continuous time) or recursive (if in discrete time) equations. It provides a powerful and general mathematical language and toolbox for examining phenomena in such systems which are generic, that is, independent from their specific physical realization, and that exist across timescales. These phenomena include, for instance, oscillations, synchronization among units of a system, attractor states, phase transitions, or deterministic chaos. Although generic and formulated in an abstract language, these phenomena are not merely conceptual or even metaphorical, but 'real'. They are experimentally and clinically accessible and quantifiable processes that can be *measured* and *inferred* from data, and that *determine* and *predict* future developments and prescribe how to best influence the system. As such, they should hence have a prominent place in guiding interventions.

As we will argue in this article, DST may serve as a kind of hub, a central layer of convergence or level of nervous system description at which phenomena relevant to mental illness could be understood, explained, classified and predicted. DST represents a layer of convergence in the sense that a number of very different, seemingly unrelated physiological and anatomical processes may give rise to similar alterations in network dynamics and behavior (Fig. 1). This may explain why quite different causal factors and pathogenic routes may give rise to similar phenomena [c.f. (3)]. At the same time, the same changes in network dynamics may be associated with a variety of quite different symptoms (Fig. 1), depending on the brain areas in which these dynamical alterations are mostly expressed. This emphasis on the *dynamical systems level* also bears important implications for the

treatment of mental illness, as discussed in sect. 3.

The idea of this article is to introduce important DST concepts and phenomena directly within the context of neuroscientific and psychiatric observations they may account for [see also (4-6)], and to illustrate them based on the same formal example of a dynamical system (DS), a recurrent neural network (RNN) model (Fig. 3A), with more formal background included in the Supplement. We will also briefly address how DS can be inferred from observations.

## 2. Dynamical phenomena and their potential relation to psychiatric conditions

A DS is described by a set of system variables (like membrane potentials or symptom strengths) and equations governing their temporal evolution (see DST primer in Supplement). A comprehensive geometrical representation of a DS is its state space, which is the space spanned by all its dynamical variables, as illustrated in Fig. 2A. A nice and powerful property of the state space representation is that it provides a complete description of the system's state, behavior, and (in the deterministic case) future fate: A point in this space exhaustively specifies the system's current state (i.e., the current values of all variables describing the system), and the so-called flow (vector) field (the arrows in Fig. 2B) completely specifies how the system will evolve in time when released at any point in this space (namely, along the direction indicated by the vectors). The temporal evolution of the system's state within this space when started from a specific initial condition is represented by its trajectory (Fig. 2). In essence, the system’s trajectory in state space shows how all variables jointly evolve in time; there is a 1:1 correspondence between such a trajectory and the more familiar time graphs of all variables (Fig. 2A).

Consider as a very simple psychological example the interaction between psychological stress, mood, and social retreat, as depicted in Fig. 2A. As stress levels increase, with some delay mood will decline, which in turn may lead to an increased tendency to retreat from the world and social interactions. As a consequence stress levels may drop again, mood will tend to increase, and the person may increasingly engage again in social and job-related responsibilities, potentially starting the whole cycle all over again. Such cyclic interactions between variables are commonly observed in the setting of mental health and are important tools for instance in case formulations in psychotherapy. Indeed, interactions between symptoms have been argued to characterize the long-term course of illnesses better than standard latent-factor models (3, 7).

## 2.1. Attractor dynamics and multi-stability

Fig. 3B illustrates the flow field for a simple 2-dimensional formal example of a DS, a 2-unit RNN (Fig. 3A). The flow field indicates a specific geometry of the state space that determines the fate of trajectories when released at specific initial conditions: In this case, the state space contains three ***fixed points***, points at which the flow becomes exactly zero in all directions (i.e., the vectors vanish). Two of them are *stable* (solid dots) in the sense that activity converges to them from all directions, hence small displacements (*perturbations*) decay back to them. Such stable fixed points are also called *point attractors*, and they are surrounded by a *basin of attraction* which is the set of all points from which activity converges to the respective point attractor. The fixed point in the center (open dot), in contrast, is *unstable* with activity diverging along at least one direction (fixed points with both converging and diverging directions are called *saddle points*). If **noise is added** to the dynamical system, it may cause trajectories to eventually cross the 'energy ridge' between attractors (Fig. 3E). The likelihood of such transitions or, conversely, the dwell times within specific states, will depend on the noise amplitude and the steepness of the attractor basins, i.e. the magnitude of the opposing flow (Fig. 3C,D). This gives rise to a phenomenon called 'meta-stability' (8), where noise-induced perturbations can cause the system to hop around different attractor states (Fig. 3C).

Dynamical systems may harbor many different stable fixed points, or other attractor objects. Such multi-stability, that is the co-existence of many attractor states, has been proposed to underlie functions like working memory (9, 10), with each fixed point corresponding to the active maintenance of a different memory item. The idea is that different briefly-presented stimuli would push the network into one of the different stimulus-specific attractor states (Fig. 3B), which – by virtue of their attractor property – would maintain an active representation of the stimulus even after its removal. Physiologically, the attractor states may correspond to elevated firing rates (termed 'persistent activity') in the respective subset of stimulus-selective neurons [e.g., (11, 12)], and may be established through strong recurrent excitation within this subpopulation (or 'cell assembly'; (9, 13)). Attractor dynamics are thought to play a role also in a variety of other cognitive processes, including decision making (14-18), probabilistic (Bayesian) inference (19), the formation and maintenance of beliefs (20, 21), or processes like memory recollection and pattern completion (19, 22-24). Formally, models of reinforcement learning are DS as well (25, 26) that may settle into stable fixed points in a stationary environment.

Profound alterations in attractor dynamics may impact mental functions. As a biophysical-level example, dopamine via its synaptic and ionic actions can regulate the width and steepness of basins of attraction (Fig. 3E), with the direction of change depending on the receptor subtype (D1- vs. D2-class) primarily stimulated (27-33). This could alter the tradeoff between cognitive flexibility, supported by flat attractor basins that ease moving among representations, vs. working memory and goal

orientation, facilitated by deep and wide basins that protect the current focus (27, 34). Via these dynamical mechanisms, the changes in the dopaminergic system known in schizophrenia may therefore account for the observed deficits in both working memory and cognitive flexibility (27, 35, 36).

Consider as another example impaired emotion regulation in depression. Ramirez-Mahaluf and Compte (37)viewed this as emerging from the mutually inhibitory interaction between an 'emotional' and a 'cognitive' hub, namely the ventral anterior cingulate cortex (vACC) and the dorsolateral prefrontal cortex (dlPFC), respectively. According to their model, high glutamatergic tone in the vACC results in overly stable attractor states which inhibit 'cognitive' activation in the dlPFC. Within a certain parameter regime, this could be counteracted by serotonine-induced hyperpolarization of vACC neurons through SSRIs. This idea is illustrated in Fig. 3B-D with two units (which one may think of as representing two network hubs in this context) with strong self-excitation but mutual inhibition (i.e., negative weights $w_{12}$ and $w_{21}$ and positive weights $w_{11}$ and $w_{22}$ in Fig. 3A). As illustrated, by either increasing the amount of self-excitation in one of the two hubs or through an imbalance in the feedback between the two (Fig. 3D), one of the two attractor basins may strongly expand at the expense of the other. At the psychological level, this type of account would also explain why strong rumination and negative mood (reflecting a strong emotion attractor) concur with a lack of attention and impaired decision-making (38, 39), or why increased fear may inhibit performance under high cognitive load and vice versa (40).

As in the example of working memory, strong attractor states often arise through positive feedback loops. For instance, stressful life events predict depression, but are also caused by depression (41), raising the possibility that after a first life event, further life events may be caused by the depressed state, leading to a mutually reinforcing feedback loop between stress and depression. Conflict states in couples (42) and groups (43) may manifest through similar attractor dynamics, with positive feedback loops leading to escalation, emphasizing the role of de-escalation techniques. While space constraints prevent a more detailed discussion, we point to similar studies highlighting the role of attractor dynamics in the domains of ketamine (32, 44), dopamine and schizophrenia (27-32, 45), depression (46), attention-deficit hyperactivity disorder (47-51), obsessive-compulsive disorder (52, 53), and post-traumatic stress disorder (54-56) (see also Table 1).

## 2.2. Sequential phenomena: limit cycles and heteroclinic channels

Fig. 4A illustrates another setup of the RNN. A slight change in some of the system parameters (cf. Supplement) gives rise to a different set of phenomena: Rather than converging to a stable fixed point, the RNN now continues to periodically oscillate. It is not a simple harmonic (sinusoidal-type) oscillation, however, but a more complex waveform that is repetitively produced. This complex but

still periodic waveform represents another type of attractor state, termed a stable *limit cycle* (just as with fixed points, there are also unstable limit cycles). Limit cycles can become quite complicated in appearance, with multiple different minima and maxima and very long periods in which the system's trajectory does not precisely retrace itself, although they always close up eventually. Limit cycles often result from interacting *positive and negative feedback loops*, as ubiquitous in the nervous system. They may represent sequences that are to be repetitively produced, like potentially complex, but still relatively stereotypical motor programs or movement patterns (57-60).

Stereotypical, repeating movement patterns are observed in many neurological and some psychiatric conditions (61, 62). In general, nonlinear oscillations – the equivalent of limit cycles in the time domain – are a hallmark of nervous system activity (63), and specific alterations for instance in the gamma or delta frequency band have indeed been described in schizophrenia (64) or ADHD (65). At a higher cognitive level, perseveratively reoccurring chains of the same thoughts may potentially be generated this way. Stable limit cycles may also underlie many types of symptom clusters which emerge in periodic intervals (66-68).

There are also other, more flexible ways to generate sequences in DS, as illustrated in Fig. 4B (69, 70). Here, orbits connect a chain of saddle points, that is, 'half-stable' fixed points towards which activity converges from some directions but leaves along others. The curves that connect the different saddle points are called 'heteroclinic orbits' (Fig. 4B), and the whole arrangement of heteroclinic orbits connecting a chain of saddle points has been termed a 'heteroclinic channel' [HC; (69, 70)]. The HC acts like an attractor, pulling in trajectories from the vicinity which then, with a bit of noise, travel along the curves connecting the saddle points which may implement a sequence of thoughts or actions. Unlike a limit cycle, the HC is not necessarily automatically repeating – it may start and terminate in a stable fixed point (as in the example in Fig. 4B). More importantly, this arrangement is much more flexible: While limit cycles determine a rather rigid sequence of events, in a HC saddle points could more easily be removed from or added to the already existing sequence through proper parameter changes, making this a more plausible account for higher cognitive functions (e.g., syntactical sequences) than limit cycles (69). Indeed, it has been suggested that belief sets evolve as heteroclinic channels in the course of psychotherapy (71).

Whether psychiatric phenomena with 'periodic' behavior, e.g. alternation between relapse and remission episodes, can be better described in terms of limit cycles, HC, or hopping among meta-stable states, is a difficult but, in principle, empirically tractable question. Differentiating among these scenarios could have important implications for optimal treatment, both in terms of the type and the *timing* of an intervention (72).

## 2.3. Attractor ghosts and the regulation of flow

Another important phenomenon in dynamical systems is that of 'attractor ruins' (73, 74), also termed 'attractor ghosts' (75), quasi- or semi-attracting states (76). These are 'attractors' which are *almost* stable, i.e. to which trajectories still converge along most directions but may slowly escape along others (Fig. 5B,C). In these scenarios, the system's parameters are *very close* to a configuration which would yield a true attractor, just not quite there (Fig. 6).

This comes with important and interesting implications that differentiate these objects from either true attractors or clearly unstable objects. Imagine a scenario where attractor valleys (Fig. 3E) become perfectly flat along one or more directions. This gives rise to a so-called **line-attractor** where the fixed points form a line, ring or plane (77-80), a continuum of *neutrally* stable fixed points along which there is neither con- nor divergence (Fig. 5A). Line attractors have been proposed to underlie phenomena such as parametric working memory (78) where a continuously valued quantity [like a 'flutter frequency' (81) or spatial position (82)] has to be retained in working memory. An attractor ruin results, for instance, if we now slightly 'detune' the line attractor (Fig. 5B). This leads to new effective time constants which are largely independent from the system's intrinsic (e.g., biophysical) time constants, a phenomenon that has been exploited for interval timing in neural systems (77, 83). Too wide detuning may in turn account for timing problems, specifically a speedup of the internal clock, evident in Parkinson's disease (84) or ADHD (85), given that the dopaminergic system has been linked to alterations in (interval) time perception and production (86, 87). Too narrow tuning, on the other hand, could produce a slowing down of time perception as in bipolar patients (88).

Hence, trajectories considerably slow down and tend to prevail in attractor ruins. Just as with HC, this phenomenon could also be exploited for flexible sequence generation with trajectories traveling among attractor ruins (89). In consequence, alterations in attractor ruins may cause characteristic symptoms, e.g. slowed-down mental processing as often observed in MDD patients (90, 91).

## 2.4. Chaos

Chaos is a strange phenomenon where a deterministic DS exhibits *aperiodic* and *irregular* behavior even in the absence of noise, with the system's states never quite repeating themselves [Fig. 4C; (75, 92)]. The state of chaos can still be an attractor, pulling in trajectories from a larger basin of attraction into a bounded region of state space within which trajectories would continue to travel forever, yet would not form a closed orbit [i.e., limit cycle; (75)]. Unlike fixed point and limit cycle attractors, chaotic attractors have at least one direction along which trajectories *diverge*, yet get 're-injected' into the same volume of state space (75). Because of this divergence, activity on the chaotic attractor is highly (exponentially) sensitive to perturbations and minimal differences in initial conditions (Fig. 4C bottom), the famous 'butterfly effect' (93).

The fact that activity in chaotic attractors is irregular yet not random, retaining a certain sequential structure, may also be beneficial for certain cognitive and coding purposes (94). In a sense, it creates deterministic variation around a central theme which may be relevant to cognitive search and creativity (95, 96). Especially interesting from a computational perspective is the phenomenon of *chaotic itinerancy* (73, 74) where trajectories chaotically traverse between different attractor ruins (see 2.3), a setup that has been exploited for dynamic and flexible sequence production and recognition (89).

Somewhat surprisingly, placing neural systems at the edge of chaos, or slightly within a chaotic regime (97-100), has important computational benefits: Here, the system naturally expresses complex temporal structure while at the same time hanging on to external stimulus information for a while. In contrast, if the system is too regular (too convergent) it exhibits no interesting internal behavior, while if it is too chaotic (too divergent) it quickly forgets about external stimuli. Consequently, if the brain leaves this computationally optimal regime and migrates either too much into the regular or too much into the chaotic range, problems may ensue. Indeed, PTSD patients show a highly reduced heart rate variability (i.e., larger regularity), assumed to be indicative of a reduced ability to flexibly respond to incoming information (101, 102). Diminished variability in mental states has also been described in higher age (103). On the other hand, a highly chaotic regime with its sensitivity to perturbations may account for attentional problems and a high distractibility by external stimuli, as, e.g., observed in ADHD (49, 104).

As another example, some authors have argued that the seemingly random patterns of thought observed in schizophrenia, reflected in associative hopping and disorganized cognition, may be rooted in too chaotic system dynamics (see e.g., (105)). In line with this idea, a number of studies observed signatures of increased chaoticity in schizophrenia patients in electrophysiological and electrodermal recordings (106-109). Mood variations in bipolar disorder have also been characterized as increasingly chaotic patterns (110, 111), potentially driven by stronger interactions among negative affective states [(112); in general, increased coupling among network elements can lead into a chaotic regime (113)].

## 2.5. Phase transitions and bifurcations

In the discussions above we have repeatedly highlighted that many dynamical phenomena may be obtained within the very same system (Figs. 3-5), just by changing some of its parameters. This gives rise to another highly important observation: As system parameters are smoothly changed, we may encounter dramatic and abrupt, *qualitative* changes in the system's behavior at some critical point (Fig. 6)! These are points in parameter space, called *bifurcation* points, where the set of dynamical objects and/or their properties change, i.e. where certain fixed points, limit cycles or chaotic objects may come into existence, vanish, or change stability.

This observation in DS is likely to have profound implications for our understanding of crucial transitions, sudden onsets or offsets, or different distinct phases in psychiatric illnesses. That neural systems may undergo critical bifurcations with dramatic consequences is comparatively well established in epilepsy (114, 115), where one has relatively clear electrophysiological signatures that allow to identify and distinguish different types of bifurcations [see also (94, 114, 116-120)].

At a more cognitive level, bifurcations may account for sudden transitions observed in behavioral choices and the accompanying neural activity during the learning of a new rule (121). Also, both brief amnestic periods (122), during which stored memories cannot be recalled, as well as so-called lucid moments in dementia (123), where suddenly mnemonic details can be recovered again, suggest that neural systems may sometimes hover at the edge of a bifurcation. Bifurcations may also help to explain why psychopharmacological treatment sometimes helps and in other instances completely fails: Ramirez-Mahaluf, Roxin (124), for instance, mimicked the effects of increased glutamate reuptake and selective serotonin reuptake inhibitors (SSRI) on network activity, and found that while within a certain range 'healthy' attractor dynamics could be pharmacologically restored, especially after passing critical bifurcation points network changes appeared irreversible by pharmacological means. In such cases, to kick a neural system out of particularly deep attractor states, more profound perturbations (potentially provided by interventions such as electro-convulsive therapy (ECT) or deep-brain stimulation (DBT)) may be required (125, 126). Profound changes reminiscent of crossing bifurcation points have also been proposed as explanation for therapy resistance in schizophrenia (72). The change between states of depression and health also shows signatures that are typical of a phase transition, namely so-called critical slowing down (similar as in Fig. 5B) where the autocorrelation length of different emotions increases (127).

Hence, from a DS perspective, one may see therapeutic efforts in psychiatry as attempts to prevent certain bifurcations from happening or to induce others.

## 2.6. Inferring DST phenomena from empirical observations

DST will of course only be useful to the clinic if the discussed phenomena can be measured and characterized. Correlation- or choherence-based analyses (128-130), power spectra (131-133), or tools like Dynamic Causal Modeling (DCM; (134-137)), have been used for some time to characterize changes in functional connectivity among brain nodes and other temporal properties of neurophysiological time series. However, almost all of these tools are *linear*, or, like Hidden Markov Models (HMM; (138, 139)), come with strong assumptions and restrictions. Linear models *cannot* produce most of the dynamical systems phenomena discussed here, except for simple phenomena like isolated fixed points, line attractors, or simple (unstable/ neutrally stable) harmonic oscillations. They

are therefore not suitable for addressing DST phenomena more generally (see, e.g., (140)). Some nonlinear aspects of the dynamics can be inferred from scaling laws [(141), but see (142)], perturbation approaches (143, 144), change points (121, 145), delay embeddings (146), or other properties of the observed time series (23, 147). While such methods provide important signatures of specific phenomena (e.g., a bifurcation), they do not return a full picture of the system dynamics.

More recently, however, progress in machine learning made it possible to extract attractor dynamics directly from empirical time series such as EEG or fMRI measurements (140), or ecological momentary assessments (EMA; (148)), using generic nonlinear dynamical systems formulations set up to approximate whatever set of unknown governing equations may have generated the empirical observations (140, 149-151). RNN are particularly well-suited for this purpose: there are mathematical theorems that assure us that RNN are *dynamically universal* in the sense that (almost) any other DS can be reformulated as a dynamically equivalent RNN that will produce the same flow field and thus dynamics in state space (151-153). Coupled with sophisticated statistical inference and deep learning methods (140, 149), RNN can be trained to reproduce and forecast experimental time series, and ultimately to recover the underlying dynamical system itself (140, 148). For a more in depth discussion of these new developments, other methods (146, 154, 155), and some of the current limitations and caveats, we refer the reader to the Supplement and to (140).

## 3. Implications: dynamics as treatment targets

Computational system dynamics provides an inherently *translational* language that could be used to describe diverse phenomena at biophysical, cognitive and even societal levels in the very same DS terms, in the language of state spaces, trajectories and attractors (69, 77, 116, 156). It thereby enables, for instance, findings in animals to be directly related to findings in humans, or to transfer mechanistic DS insights from one physiological or behavioral domain to another. Approaches for reconstructing DS from data (140, 149, 157-160) are even relatively agnostic to the precise measurement modality (except for limitations from a method's temporal or spatial resolution), that is the *same* DS may be inferred from neuroimaging, surface electrode, multiple single-unit recordings, or behavioral data.

The most critical contribution of DST is likely an appreciation of dynamical rather than static features as potential targets of interventions. Consider an example from schizophrenia. Traditionally, the focus has been to use medication to directly reverse known physiological aberrations, e.g. through dopaminergic antagonists. However, DS are complex, and many diverse ionic effects may act *synergistically* to establish a certain dynamical regime (161) (28). Restoring only *part* of the ionic functions underlying the original deficits, e.g., *only* GABAergic transmission, could – counterintuitively – make the situation even *worse* (9). On the other hand, from a DS perspective,

pharmacological agents may not have to target exactly those transmitter systems most compromised in a disease. Perhaps it is easier, cheaper, or more biocompatible to use compounds which target alternative mechanisms, for instance cellular calcium channels, with exactly the same implications for dynamics that dopaminergic drugs would have. Because there are usually many different and mutually redundant routes to the same dynamical phenomena, very different treatments could have similar effects. Moreover, some dynamic feedback loops in the brain may be much more sensitive to parameter changes than others, rendering them more effective targets for treatment than others. Appreciation of dynamical properties may hence open up new paths for intervention.

One other implication is that assessing and monitoring the system dynamics in patients or at-risk subjects may be more informative than the current approach of examining subjective phenomena by asking individuals to average over periods ranging from weeks to months (thus averaging out temporal dynamics). The DST perspective may also help us to understand how seemingly unrelated phenomena are truly connected at a deeper level (leading, e.g., into comorbidities): Perhaps the brain becomes generally vulnerable to a specific type of alteration of its dynamical regimes (e.g., due to a transmitter imbalance) – depending on which brain areas are affected most by these dynamical alterations, they will find their incarnation in different bundles of symptoms (see Fig. 1). For instance, while hyper-stable attractor states in auditory areas may cause tinnitus, the same alterations in orbitofrontal cortex may be associated with perseveration of suboptimal responses.

In conclusion, appreciating the dynamical properties of mental illnesses could have profound implications for how we diagnose, classify, predict, and treat psychiatric symptoms. Currently, however, this field is still very much in its infancy, and much more systematic empirical studies that directly address DST mechanisms in psychiatry are certainly needed, potentially building on new methodological developments in the fields of machine and deep learning (sect. 2.6).

## Acknowledgments and Disclosures

Dr. Durstewitz received funding from the German Science Foundation (DFG; Du 354/8-2, Du 354/10-1) and from the German Federal Ministry of Education and Research (BMBF) within the e:Med program (01ZX1311A [SP7] & 01ZX1314G [SP10]). Dr. Koppe received funding from the German Science Foundation (DFG; TRR265: A06 & B08). Dr. Huys acknowledges support by the UCLH NIHR BRC.

The authors report no biomedical financial interests or potential conflicts of interest.

An earlier version of this manuscript has been posted on arXiv:1809.06303.

# 4. References

1. Bystritsky A, Nierenberg A, Feusner J, Rabinovich M. Computational non-linear dynamical psychiatry: a new methodological paradigm for diagnosis and course of illness. Journal of psychiatric research. 2012;46(4):428-35.

2. Molenaar PC, Campbell CG. The new person-specific paradigm in psychology. Current directions in psychological science. 2009;18(2):112-7.

3. Kendler KS, Zachar P, Craver C. What kinds of things are psychiatric disorders? Psychological medicine. 2011;41(6):1143-50.

4. Roberts JA, Friston KJ, Breakspear M. Clinical Applications of Stochastic Dynamic Models of the Brain, Part II: A Review. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2017;2(3):225-34.

5. Breakspear M. Dynamic models of large-scale brain activity. Nature neuroscience. 2017;20:340.

6. Wang XJ, Krystal JH. Computational psychiatry. Neuron. 2014;84(3):638-54.

7. Borsboom D, Cramer AO, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. PloS One. 2011;6(11):e27407.

8. Balaguer-Ballester E, Moreno-Bote R, Deco G, Durstewitz D. Metastable dynamics of neural ensembles. Frontiers in Systems Neuroscience. 2017;11:99.

9. Durstewitz D, Seamans JK, Sejnowski TJ. Neurocomputational models of working memory. Nature neuroscience. 2000;3 1184-91.

10. Wang XJ. Synaptic reverberation underlying mnemonic persistent activity. Trends in neurosciences. 2001;24(8):455-63.

11. Funahashi S, Bruce CJ, Goldman-Rakic PS. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. Journal of Neurophysiology. 1989;61(2):331-49.

12. Fuster J. The prefrontal cortex: Academic Press; 2015.

13. Hebb DO, Hebb D. The organization of behavior: Wiley New York; 1949.

14. Wang X-J. Decision making in recurrent neuronal circuits. Neuron. 2008;60(2):215-34.

15. Wang X-J. Probabilistic decision making by slow reverberation in cortical circuits. Neuron. 2002;36(5):955-68.

16. Albantakis L, Deco G. The encoding of alternatives in multiple-choice decision-making. BMC Neuroscience. 2009;10(1):166.

17. Ratcliff R, McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural Computation. 2008;20(4):873-922.

18. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review. 2006;113(4):700.

19. Lengyel M, Kwag J, Paulsen O, Dayan P. Matching storage and recall: hippocampal spike timing–dependent plasticity and phase response curves. Nature neuroscience. 2005;8(12):1677.

20. Heskes T, editor Stable fixed points of loopy belief propagation are local minima of the bethe free energy. Advances in neural information processing systems; 2003.

21. Adams RA, Napier G, Roiser JP, Mathys C, Gilleen J. Attractor-like dynamics in belief updating in schizophrenia. Journal of Neuroscience. 2018;38(44):9471-85.

22. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences U S A. 1982;79(8):2554-8.

23. Wills TJ, Lever C, Cacucci F, Burgess N, O'keefe J. Attractor dynamics in the hippocampal representation of the local environment. Science. 2005;308(5723):873-6.

24. Ratcliff R. A theory of memory retrieval. Psychological Review. 1978;85(2):59.

25. Durstewitz D. Advanced data analysis in neuroscience: Springer; 2017.

26. Gershman SJ. A unifying probabilistic view of associative learning. PLoS computational biology. 2015;11(11):e1004567.

27. Durstewitz D, Seamans JK. The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. Biological Psychiatry. 2008;64(9):739-49.

28. Durstewitz D, Seamans JK, Sejnowski TJ. Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. Journal of Neurophysiology. 2000;83(3):1733-50.

29. Lapish CC, Balaguer-Ballester E, Seamans JK, Phillips aG, Durstewitz D. Amphetamine Exerts Dose-Dependent Changes in Prefrontal Cortex Attractor Dynamics during Working Memory. Journal of Neuroscience. 2015;35(28):10172-87.

30. Gruber AJ, Solla SA, Surmeier DJ, Houk JC. Modulation of striatal single units by expected reward: a spiny neuron model displaying dopamine-induced bistability. Journal of Neurophysiology. 2003;90(2):1095-114.

31. Maia TV, Cano-Colino M. The role of serotonin in orbitofrontal function and obsessive-compulsive disorder. Clinical Psychological Science. 2015;3(3):460-82.

32. Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, et al. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. Cerebral cortex (New York, NY : 1991). 2014;24(4):859-72.

33. King R, Barchas JD, Huberman B, editors. Theoretical Psychopathology: An Application of Dynamical Systems Theory to Human Behavior. Synergetics of the Brain; 1983 1983//; Berlin, Heidelberg: Springer Berlin Heidelberg.

34. Ueltzhöffer K, Armbruster-Genç DJ, Fiebach CJ. Stochastic dynamics underlying cognitive stability and flexibility. PLoS Computational Biology. 2015;11(6):e1004331.

35. Armbruster DJ, Ueltzhöffer K, Basten U, Fiebach CJ. Prefrontal cortical mechanisms underlying individual differences in cognitive flexibility and stability. Journal of Cognitive Neuroscience. 2012;24(12):2385-99.

36. Floresco SB, Block AE, Maric T. Inactivation of the medial prefrontal cortex of the rat impairs strategy set-shifting, but not reversal learning, using a novel, automated procedure. Behavioural Brain Research. 2008;190(1):85-96.

37. Ramirez-Mahaluf JP, Compte A. Serotonergic Modulation of Cognition in Prefrontal Cortical Circuits in Major Depression. Computational Psychiatry: Elsevier; 2018. p. 27-46.

38. Gotlib IH, Joormann J. Cognition and Depression: Current Status and Future Directions. Annual Review of Clinical Psychology. 2010;6:285-312.

39. Lyubomirsky S, Kasri F, Zehm K. Dysphoric rumination impairs concentration on academic tasks. Cognitive Therapy and Research. 2003;27(3):309-30.

40. Vytal K, Cornwell B, Arkin N, Grillon C. Describing the interplay between anxiety and cognition: from impaired performance under low cognitive load to reduced anxiety under high load. Psychophysiology. 2012;49(6):842-52.

41. Kendler KS, Karkowski LM, Prescott CA. Causal relationship between stressful life events and the onset of major depression. American Journal of Psychiatry. 1999;156(6):837-41.

42. Gottman J, Swanson C, Murray J. The mathematics of marital conflict: Dynamic mathematical nonlinear modeling of newlywed marital interaction. Journal of Family Psychology. 1999;13(1):3.

43. Coleman PT, Vallacher RR, Nowak A, Bui-Wrzosinska L. Intractable conflict as an attractor: A dynamical systems approach to conflict escalation and intractability. American Behavioral Scientist. 2007;50(11):1454-75.

44. Starc M, Murray JD, Santamauro N, Savic A, Diehl C, Cho YT, et al. Schizophrenia is associated with a pattern of spatial working memory deficits consistent with cortical disinhibition. Schizophrenia research. 2017;181:107-16.

45. Braun U, Harneit A, Pergola G, Menara T, Schaefer A, Betzel RF, et al. Brain state stability during working memory is explained by network control theory, modulated by dopamine D1/D2 receptor function, and diminished in schizophrenia. arXiv preprint arXiv:190609290. 2019.

46. Cramer AO, van Borkulo CD, Giltay EJ, van der Maas HL, Kendler KS, Scheffer M, et al. Major

depression as a complex dynamic system. PloS one. 2016;11(12):e0167490.

47. Forster S, Lavie N. Establishing the Attention-Distractibility Trait. Psychological science. 2016;27(2):203-12.

48. Wilens TE, Faraone SV, Biederman J. Attention-deficit/hyperactivity disorder in adults. Journal of the American Medical Association. 2004;292(5):619-23.

49. Bubl E, Dorr M, Riedel A, Ebert D, Philipsen A, Bach M, et al. Elevated background noise in adult attention deficit hyperactivity disorder is associated with inattention. PLoS One. 2015;10(2):e0118271.

50. Cortese S, Kelly C, Chabernaud C, Proal E, Di Martino A, Milham MP, et al. Toward systems neuroscience of ADHD: a meta-analysis of 55 fMRI studies. American Journal of Psychiatry. 2012;169(10):1038-55.

51. Hauser TU, Fiore VG, Moutoussis M, Dolan RJ. Computational Psychiatry of ADHD: Neural Gain Impairments across Marrian Levels of Analysis. Trends in neurosciences. 2016;39(2):63-73.

52. Rolls ET, Loh M, Deco G. An attractor hypothesis of obsessive–compulsive disorder. The European journal of neuroscience. 2008;28(4):782-93.

53. Rabinovich MI, Muezzinoglu MK, Strigo I, Bystritsky A. Dynamical principles of emotion-cognition interaction: mathematical images of mental disorders. PloS one. 2010;5(9):e12547.

54. Lanius RA, Brand B, Vermetten E, Frewen PA, Spiegel D. The dissociative subtype of posttraumatic stress disorder: Rationale, clinical and neurobiological evidence, and implications. Depression and Anxiety. 2012;29(8):701-8.

55. Lanius RA, Vermetten E, Loewenstein RJ, Brand B, Schmahl C, Bremner JD, et al. Emotion modulation in PTSD: Clinical and neurobiological evidence for a dissociative subtype. American Journal of Psychiatry. 2010;167(6):640-7.

56. Sack M, Cillien M, Hopper JW. Acute dissociation and cardiac reactivity to script-driven imagery in trauma-related disorders. European Journal of Psychotraumatology. 2012;3(1):17419.

57. Kato S, Kaplan HS, Schrodel T, Skora S, Lindsay TH, Yemini E, et al. Global brain dynamics embed the motor command sequence of Caenorhabditis elegans. Cell. 2015;163(3):656-69.

58. Marder E, Bucher D. Central pattern generators and the control of rhythmic movements. Current Biology. 2001;11(23):R986-R96.

59. Marder E, Goeritz ML, Otopalik AG. Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. Current Opinion in Neurobiology. 2015;31:156-63.

60. Russo AA, Bittner SR, Perkins SM, Seely JS, London BM, Lara AH, et al. Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. Neuron. 2018;97(4):953-66.e8.

61. Ridley R. The psychology of perseverative and stereotyped behaviour. Progress in Neurobiology. 1994;44(2):221-31.

62. Turner M. Annotation: Repetitive behaviour in autism: A review of psychological research. The Journal of Child Psychology and Psychiatry and Allied Disciplines. 1999;40(6):839-49.

63. Buzsáki G, Draguhn A. Neuronal oscillations in cortical networks. Science. 2004;304(5679):1926-9.

64. Uhlhaas PJ, Haenschel C, Nikolic D, Singer W. The role of oscillations and synchrony in cortical networks and their putative relevance for the pathophysiology of schizophrenia. Schizophrenia bulletin. 2008;34(5):927-43.

65. Demanuele C, James C, Capilla A, Sonuga-Barke E, editors. Extracting event-related field components through space-time ICA: A study of MEG recordings from children with ADHD and controls. 4th European Conference of the International Federation for Medical and Biological Engineering; 2009: Springer.

66. Chang S-S, Chou T. A Dynamical Bifurcation Model of Bipolar Disorder Based on Learned Expectation and Asymmetry in Mood Sensitivity. Comput Psychiatr. 2018;2:205-22.

67. Eldar E, Rutledge RB, Dolan RJ, Niv Y. Mood as Representation of Momentum. Trends in cognitive sciences. 2016;20(1):15-24.

68. Eldar E, Niv Y. Interaction between emotional state and learning underlies mood instability. Nature Communications. 2015;6:6149.

69. Rabinovich MI, Huerta R, Varona P, Afraimovich VS. Transient cognitive dynamics, metastability, and decision making. PLoS Computational Biology. 2008;4(5):e1000072.

70. Rabinovich MI, Varona P, Selverston AI, Abarbanel HD. Dynamical principles in neuroscience. Reviews of Modern Physics. 2006;78(4):1213.

71. Kronemyer D, Bystritsky A. A non-linear dynamical approach to belief revision in cognitive behavioral therapy. Frontiers in Computational Neuroscience. 2014;8(55).

72. Krystal JH, Anticevic A, Murray JD, Glahn D, Driesen N, Yang G, et al. Clinical Heterogeneity Arising from Categorical and Dimensional Features of the Neurobiology of Psychiatric Diagnoses. In: Redish AD, Gordon JA, editors. Computational Psychiatry: MIT Press; 2016. p. 293-316.

73. Tsuda I. Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. Behavioral and Brain Sciences. 2001;24(5):793-810.

74. Tsuda I. Chaotic itinerancy and its roles in cognitive neurodynamics. Current Opinion in Neurobiology. 2015;31:67-71.

75. Strogatz SH. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering: CRC Press; 2018.

76. Balaguer-Ballester E, Lapish CC, Seamans JK, Durstewitz D. Attracting dynamics of frontal cortex ensembles during memory-guided decision-making. PLoS Computational Biology. 2011;7(5):e1002057.

77. Durstewitz D. Self-organizing neural integrator predicts interval times through climbing activity. Journal of Neuroscience. 2003;23(12):5342-53.

78. Machens CK, Romo R, Brody CD. Flexible control of mutual inhibition: a neural model of two-interval discrimination. Science. 2005;307(5712):1121-4.

79. Seung HS. How the brain keeps the eyes still. Proceedings of the National Academy of Sciences. 1996;93(23):13339-44.

80. Seung HS, Lee DD, Reis BY, Tank DW. Stability of the memory of eye position in a recurrent network of conductance-based model neurons. Neuron. 2000;26(1):259-71.

81. Romo R, Brody CD, Hernández A, Lemus L. Neuronal correlates of parametric working memory in the prefrontal cortex. Nature. 1999;399(6735):470-3.

82. Zhang K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. Journal of Neuroscience. 1996;16(6):2112-26.

83. Durstewitz D. Neural representation of interval time. Neuroreport. 2004;15(5):745-9.

84. Rammsayer T, Classen W. Impaired temporal discrimination in Parkinson's disease: temporal processing of brief durations as an indicator of degeneration of dopaminergic neurons in the basal ganglia. International Journal of Neuroscience. 1997;91(1-2):45-55.

85. Rubia K, Halari R, Christakou A, Taylor E. Impulsiveness as a timing disturbance: neurocognitive abnormalities in attention-deficit hyperactivity disorder during temporal processes and normalization with methylphenidate. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2009;364(1525):1919-31.

86. Hass J, Durstewitz D. Neurocomputational models of time perception. Neurobiology of Interval Timing: Springer; 2014. p. 49-71.

87. Rammsayer TH. On dopaminergic modulation of temporal information processing. Biological Psychology. 1993;36(3):209-22.

88. Northoff G, Magioncalda P, Martino M, Lee HC, Tseng YC, Lane T. Too Fast or Too Slow? Time and Neuronal Variability in Bipolar Disorder-A Combined Theoretical and Empirical Investigation. Schizophr Bull. 2018;44(1):54-64.

89. Russo E, Treves A. Cortical free-association dynamics: Distinct phases of a latching network. Physical Review E. 2012;85(5):051920.

90. Tsourtos G, Thompson J, Stough C. Evidence of an early information processing speed deficit in unipolar major depression. Psychological medicine. 2002;32(2):259-65.

91. Marazziti D, Consoli G, Picchetti M, Carlini M, Faravelli L. Cognitive impairment in major depression.

European journal of pharmacology. 2010;626(1):83-6.

92. Ott E. Chaos in dynamical systems: Cambridge University Press; 2002.

93. Lorenz EN. Deterministic nonperiodic flow. Journal of the atmospheric sciences. 1963;20(2):130-41.

94. Durstewitz D, Gabriel T. Dynamical basis of irregular spiking in NMDA-driven prefrontal cortex neurons. Cerebral cortex (New York, NY : 1991). 2007;17(4):894-908.

95. Zausner T. The creative chaos: Speculations on the connection between non-linear dynamics and the creative process. Nonlinear dynamics in human behavior: World Scientific; 1996. p. 343-9.

96. Schuldberg D. Chaos theory and creativity. Encyclopedia of creativity. 1999;1:259-72.

97. Bertschinger N, Natschläger T. Real-time computation at the edge of chaos in recurrent neural networks. Neural Computation. 2004;16(7):1413-36.

98. Jaeger H, Haas H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science. 2004;304(5667):78-80.

99. Legenstein R, Maass W. Edge of chaos and prediction of computational performance for neural circuit models. Neural Networks. 2007;20(3):323-34.

100. Sussillo D, Abbott LF. Generating coherent patterns of activity from chaotic neural networks. Neuron. 2009;63(4):544-57.

101. Thayer JF, Lane RD. A model of neurovisceral integration in emotion regulation and dysregulation. Journal of affective disorders. 2000;61(3):201-16.

102. Thayer JF, Ahs F, Fredrikson M, Sollers JJ, 3rd, Wager TD. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neuroscience & Biobehavioral Reviews. 2012;36(2):747-56.

103. Battaglia D, Thomas B, Hansen EC, Chettouf S, Daffertshofer A, McIntosh AR, et al. Functional Connectivity Dynamics of the Resting State across the Human Adult Lifespan. bioRxiv. 2017.

104. Uebel H, Albrecht B, Asherson P, Börger NA, Butler L, Chen W, et al. Performance variability, impulsivity errors and the impact of incentives as gender-independent endophenotypes for ADHD. Journal of Child Psychology and Psychiatry. 2010;51(2):210-8.

105. Paulus MP, Braff DL. Chaos and schizophrenia: does the method fit the madness? Biological Psychiatry. 2003;53(1):3-11.

106. Röschke J, Mann K, Fell J. Nonlinear EEG dynamics during sleep in depression and schizophrenia. International Journal of Neuroscience. 1994;75(3-4):271-84.

107. Röschke J, Fell J, Beckmann P. Nonlinear analysis of sleep EEG data in schizophrenia: calculation of the principal Lyapunov exponent. Psychiatry research. 1995;56(3):257-69.

108. Bob P, Chladek J, Susta M, Glaslova K, Jagla F, Kukleta M. Neural chaos and schizophrenia. General physiology and biophysics. 2007;26(4):298.

109. Bob P, Susta M, Chladek J, Glaslova K, Palus M. Chaos in schizophrenia associations, reality or metaphor? International Journal of Psychophysiology. 2009;73(3):179-85.

110. Bonsall MB, Wallace-Hadrill SM, Geddes JR, Goodwin GM, Holmes EA. Nonlinear time-series approaches in characterizing mood stability and mood instability in bipolar disorder. Proceedings of the Royal Society B: Biological Sciences. 2011;279(1730):916-24.

111. Gottschalk A, Bauer MS, Whybrow PC. Evidence of chaotic mood variation in bipolar disorder. Archives of general psychiatry. 1995;52(11):947-59.

112. Wichers M, Wigman J, Myin-Germeys I. Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. Emotion Review. 2015;7(4):362-7.

113. Martignoli S, Stoop R. Phase-locking and Arnold coding in prototypical network topologies. Discrete & Continuous Dynamical Systems-B. 2008;9(1):145.

114. Naze S, Bernard C, Jirsa V. Computational modeling of seizure dynamics using coupled neuronal networks: factors shaping epileptiform activity. PLoS Computational Biology. 2015;11(5):e1004209.

115. Jirsa VK, Stacey WC, Quilichini PP, Ivanov AI, Bernard C. On the nature of seizure dynamics. Brain.

2014;137(8):2210-30.

116. Izhikevich EM. Dynamical Systems in Neuroscience: MIT Press; 2007.

117. Brunel N. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. Journal of Computational Neuroscience. 2000;8(3):183-208.

118. Durstewitz D. Implications of synaptic biophysics for recurrent network dynamics and active memory. Neural Networks. 2009;22(8):1189-200.

119. Izhikevich EM. Simple model of spiking neurons. IEEE Transactions on neural networks. 2003;14(6):1569-72.

120. Rinzel J, Ermentrout GB. Analysis of neural excitability and oscillations. Methods in Neuronal Modeling. 1998;2:251-92.

121. Durstewitz D, Vittoz NM, Floresco SB, Seamans JK. Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. Neuron. 2010;66(3):438-48.

122. Spiegel DR, Smith J, Wade RR, Cherukuru N, Ursani A, Dobruskina Y, et al. Transient global amnesia: current perspectives. Neuropsychiatric Disease and Treatment. 2017;13:2691.

123. Normann HK, Asplund K, Karlsson S, Sandman PO, Norberg A. People with severe dementia exhibit episodes of lucidity. A population-based study. Journal of Clinical Nursing. 2006;15(11):1413-7.

124. Ramirez-Mahaluf JP, Roxin A, Mayberg HS, Compte A. A Computational Model of Major Depression: the Role of Glutamate Dysfunction on Cingulo-Frontal Network Dynamics. Cerebral cortex (New York, NY : 1991). 2015;27:660-79.

125. Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C, et al. Deep brain stimulation for treatment-resistant depression. Neuron. 2005;45(5):651-60.

126. UK ECT Review Group. Efficacy and safety of electroconvulsive therapy in depressive disorders: a systematic review and meta-analysis. Lancet (London, England). 2003;361(9360):799-808.

127. van de Leemput IA, Wichers M, Cramer AO, Borsboom D, Tuerlinckx F, Kuppens P, et al. Critical slowing down as early warning for the onset and termination of depression. Proceedings of the National Academy of Sciences. 2014;111(1):87-92.

128. Sakoğlu Ü, Pearlson GD, Kiehl KA, Wang YM, Michael AM, Calhoun VD. A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia. Magnetic Resonance Materials in Physics, Biology and Medicine. 2010;23(5-6):351-66.

129. Kaiser RH, Whitfield-Gabrieli S, Dillon DG, Goer F, Beltzer M, Minkel J, et al. Dynamic resting-state functional connectivity in major depression. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2016;41(7):1822.

130. Jones DT, Vemuri P, Murphy MC, Gunter JL, Senjem ML, Machulda MM, et al. Non-stationarity in the "resting brain's" modular architecture. PloS one. 2012;7(6):e39731.

131. Uhlhaas PJ, Singer W. Abnormal neural oscillations and synchrony in schizophrenia. Nature reviews neuroscience. 2010;11(2):100.

132. Flor-Henry P, Yeudall L, Koles Z, Howarth B. Neuropsychological and power spectral EEG investigations of the obsessive-compulsive syndrome. Biological Psychiatry. 1979.

133. Grin-Yatsenko VA, Baas I, Ponomarev VA, Kropotov JD. EEG power spectra at early stages of depressive disorders. Journal of Clinical Neurophysiology. 2009;26(6):401-6.

134. Brodersen KH, Deserno L, Schlagenhauf F, Lin Z, Penny WD, Buhmann JM, et al. Dissecting psychiatric spectrum disorders by generative embedding. NeuroImage: Clinical. 2014;4:98-111.

135. Schlösser RG, Wagner G, Koch K, Dahnke R, Reichenbach JR, Sauer H. Fronto-cingulate effective connectivity in major depression: a study with fMRI and dynamic causal modeling. NeuroImage. 2008;43(3):645-55.

136. Schlösser RG, Wagner G, Schachtzabel C, Peikert G, Koch K, Reichenbach JR, et al. Fronto-cingulate effective connectivity in obsessive compulsive disorder: A study with fMRI and dynamic causal modeling. Human brain mapping. 2010;31(12):1834-50.

137. Deserno L, Sterzer P, Wüstenberg T, Heinz A, Schlagenhauf F. Reduced prefrontal-parietal effective

connectivity and working memory deficits in schizophrenia. Journal of Neuroscience. 2012;32(1):12-20.

138. Schlagenhauf F, Huys QJM, Deserno L, Rapp MA, Beck A, Heinze H-J, et al. Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. NeuroImage. 2014;89:171-80.

139. Demanuele C, Bähner F, Plichta MM, Kirsch P, Tost H, Meyer-Lindenberg A, et al. A statistical approach for segregating cognitive task stages from multivariate fMRI BOLD time series. Frontiers in human neuroscience. 2015;9:537.

140. Koppe G, Toutounji H, Kirsch P, Lis S, Durstewitz D. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. PLOS Computational Biology. 2019;15(8):e1007263.

141. Jensen HJ. Self-organized criticality: emergent complex behavior in physical and biological systems: Cambridge University Press; 1998.

142. Nonnenmacher M, Behrens C, Berens P, Bethge M, Macke JH. Signatures of criticality arise from random subsampling in simple population models. PLoS Computational Biology. 2017;13(10):e1005718.

143. Aksay E, Gamkrelidze G, Seung H, Baker R, Tank D. In vivo intracellular recording and perturbation of persistent activity in a neural integrator. Nature neuroscience. 2001;4(2):184-93.

144. Inagaki HK, Fontolan L, Romani S, Svoboda K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature. 2019;566(7743):212.

145. Toutounji H, Durstewitz D. Detecting Multiple Change Points Using Adaptive Regression Splines with Application to Neural Recordings. Frontiers in Neuroinformatics. 2018;12.

146. Takens F. Detecting strange attractors in turbulence. In: Rand DA, Young L-S, editors. Dynamical Systems and Turbulence, Lecture notes in Mathematics. 898: Springer-Verlag; 1981. p. 366-81.

147. Niessing J, Friedrich RW. Olfactory pattern classification by discrete neuronal network states. Nature. 2010;465(7294):47-52.

148. Koppe G, Guloksuz S, Reininghaus U, Durstewitz D. Recurrent Neural Networks in Mobile Sampling and Intervention. Schizophrenia bulletin. 2018;45(2):272-6.

149. Durstewitz D. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. PLoS Computational Biology. 2017;13(6):e1005542.

150. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences. 2016;113(15):3932-7.

151. Trischler AP, D'Eleuterio GM. Synthesis of recurrent neural networks for dynamical system simulation. Neural Networks. 2016;80:67-78.

152. Funahashi K-i, Nakamura Y. Approximation of dynamical systems by continuous time recurrent neural networks. Neural Networks. 1993;6(6):801-6.

153. Kimura M, Nakano R. Learning dynamical systems by recurrent neural networks from orbits. Neural Networks. 1998;11(9):1589-99.

154. Kantz H, Schreiber T. Nonlinear time series analysis: Cambridge University Press; 2004.

155. Sauer T, Yorke JA, Casdagli M. Embedology. Journal of statistical Physics. 1991;65(3):579-616.

156. Wilson HR. Spikes, decisions, and actions: the dynamical foundations of neurosciences: Oxford University Press; 1999.

157. Duncker L, Bohner G, Boussard J, Sahani M. Learning interpretable continuous-time models of latent stochastic dynamical systems. arXiv preprint arXiv:190204420. 2019.

158. Byron MY, Afshar A, Santhanam G, Ryu SI, Shenoy KV, Sahani M, editors. Extracting dynamical structure embedded in neural activity. Advances in neural information processing systems; 2006.

159. Pandarinath C, O'Shea DJ, Collins J, Jozefowicz R, Stavisky SD, Kao JC, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. Nature methods. 2018:1.

160. Roweis S, Ghahramani Z. Learning nonlinear dynamical systems using the expectation-maximization algorithm. Kalman filtering and neural networks. 2001;6:175-220.

161. Durstewitz D, Seamans JK. The computational role of dopamine D1 receptors in working memory.

Neural Networks. 2002;15(4-6):561-72.

# Tables

**Table 1.** Psychiatric symptoms and their possible dynamical systems interpretation.

| **Symptoms** | **Associated changes in attractor dynamics** |
|---|---|
| Perseveration, dissociation, obsessions, compulsions | Overly steep attractor basins |
| Distractor susceptibility/ inattentiveness, associative hopping, incoherent and disorganized thought, hallucinations | Overly flat attractor basins or increased noise levels |
| Deficits in parametric WM, jumping to conclusions (failure to integrate information) | Alterations in line attractor configurations |
| Rumination and reoccurring chains of thought, stereotypical movement patterns, persistence of invalid belief sets | Overly steep limit cycle attractors or heteroclinic channels |
| Altered time perception, slowed down mental processes | Alterations in flow around attractor ghosts |
| Lucid moments in amnesia, epileptic seizures, sudden transitions between disease stages, resistance to therapy | Bifurcations |
| Increased variability in affective states, disorganized thought, high distractibility | Too high chaoticity |
| Reduced cognitive flexibility | Too low chaoticity |

# Figure legends

**Fig. 1. Network dynamics as a layer of convergence.** A number of different physiological and structural processes (left) may give rise to similar alterations in network dynamics (center) which, depending on where in the brain they manifest, may give rise to a variety of different cognitive and emotional processes and psychiatric symptoms.

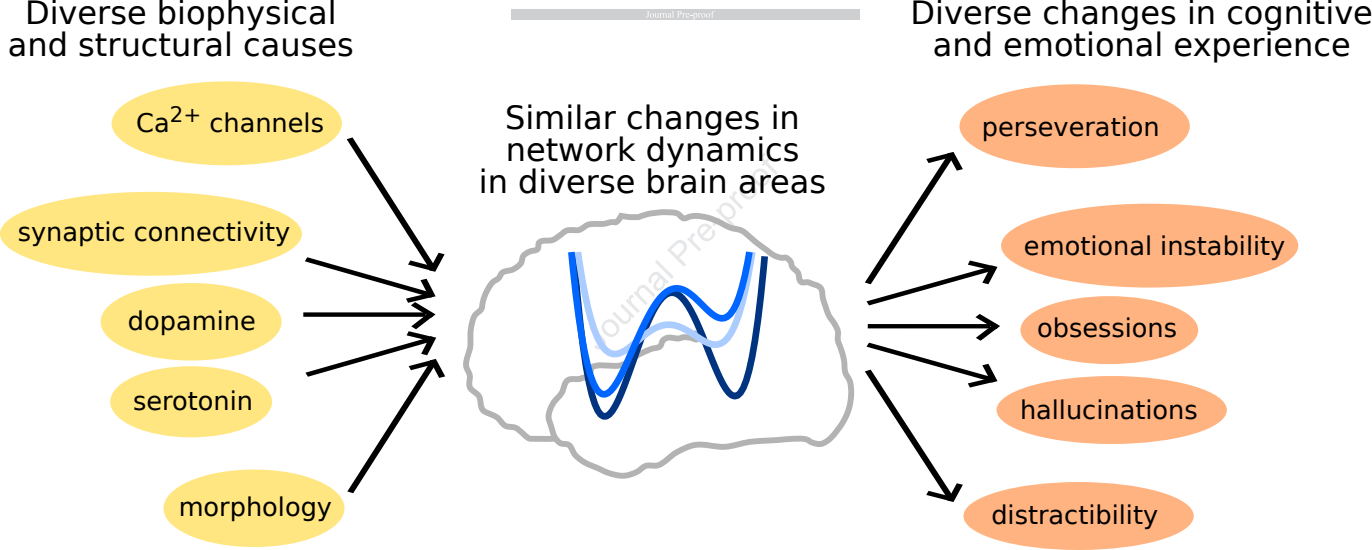
**Fig. 2. Time graphs and state spaces.** A) A central concept in DST is that of a *state space* (right). A state space is the space spanned by all dynamical variables of a system, which in this psychological example were taken to be 'mood', 'stress', and 'social retreat'. A *trajectory* in the state space corresponds to the temporal co-evolution of the dynamical variables over time, i.e. there is a 1:1 correspondence between points on the trajectory and the state of all dynamical variables when depicted as a function of time (left). Color-coding of the trajectory illustrates time progression. B) Another central concept is that of a *flow field* (right), where the vectors indicate the direction and magnitude of flow (change) at each point in state space, illustrated here with a 2-dimensional example. Examples were constructed based on the Lotka-Volterra equations.

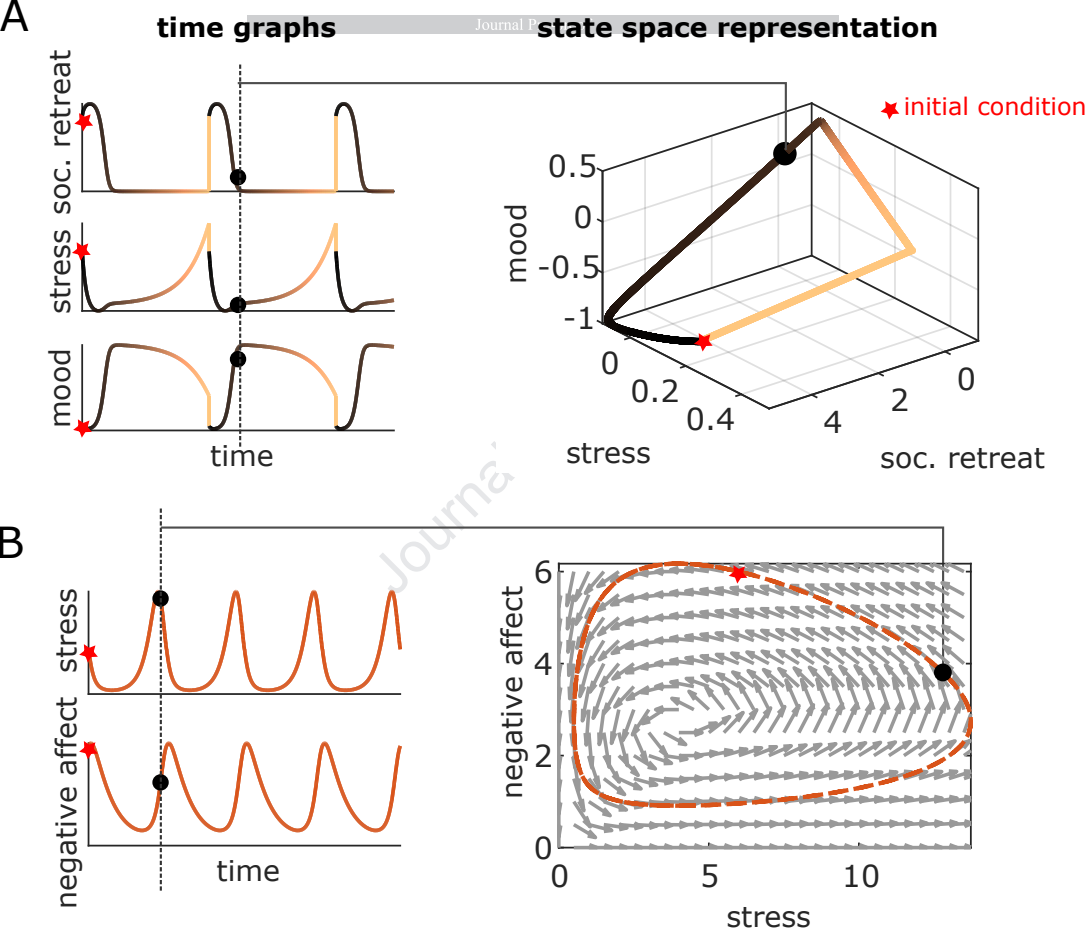
**Fig. 3. Example of multi-stability in a RNN.** A) Structure of a two-unit 'toy' RNN (see eq. 6 and parameter values used in Supplement). B) Flow field for the RNN in A, with gray arrows marking direction and magnitude of the flow. Gray-shaded lines are the so-called nullclines of unit 1 and unit 2, where the flow of either one of the two system variables vanishes, and solid/open circles show stable/unstable fixed points. The black dashed line separates the basins of attraction of the two stable fixed points. The dashed red line shows an example of a deterministic trajectory starting from the initial condition indicated by the red star (located on one of the system's two point attractors), after a brief stimulus (yellow) to unit 1. C) Same as in B) with network parameters slightly changed, causing the system's attractors to move closer together and their basins to become shallower. D) Same as in B) and C), with parameters slightly changed such that the symmetry between attractor states is broken. The system now has one attractor with steeper basin than the other. The bottom parts of figures B)-D*)* show the activation in time of unit 1, with the period of stimulation indicated in yellow. In C) and D), noise was added to the system. While in B) the network maintains unit 1's high activation by remaining in the high-rate attractor, C) shows how activity spontaneously switches between the two attractors due to the presence of noise and shallower attractor basins. In D) the system only remains briefly in the high-rate attractor due to its small basin of attraction, from which it is kicked out by the noise after relatively short dwelling times. E) Schematic potential landscape depicting the extent and depth of the basins of attraction of the systems in B) (dark grey), C) (light grey), and D) (black). Potential minima correspond to the attractor states. MATLAB code for these simulations is available at *https://github.com/DurstewitzLab.*

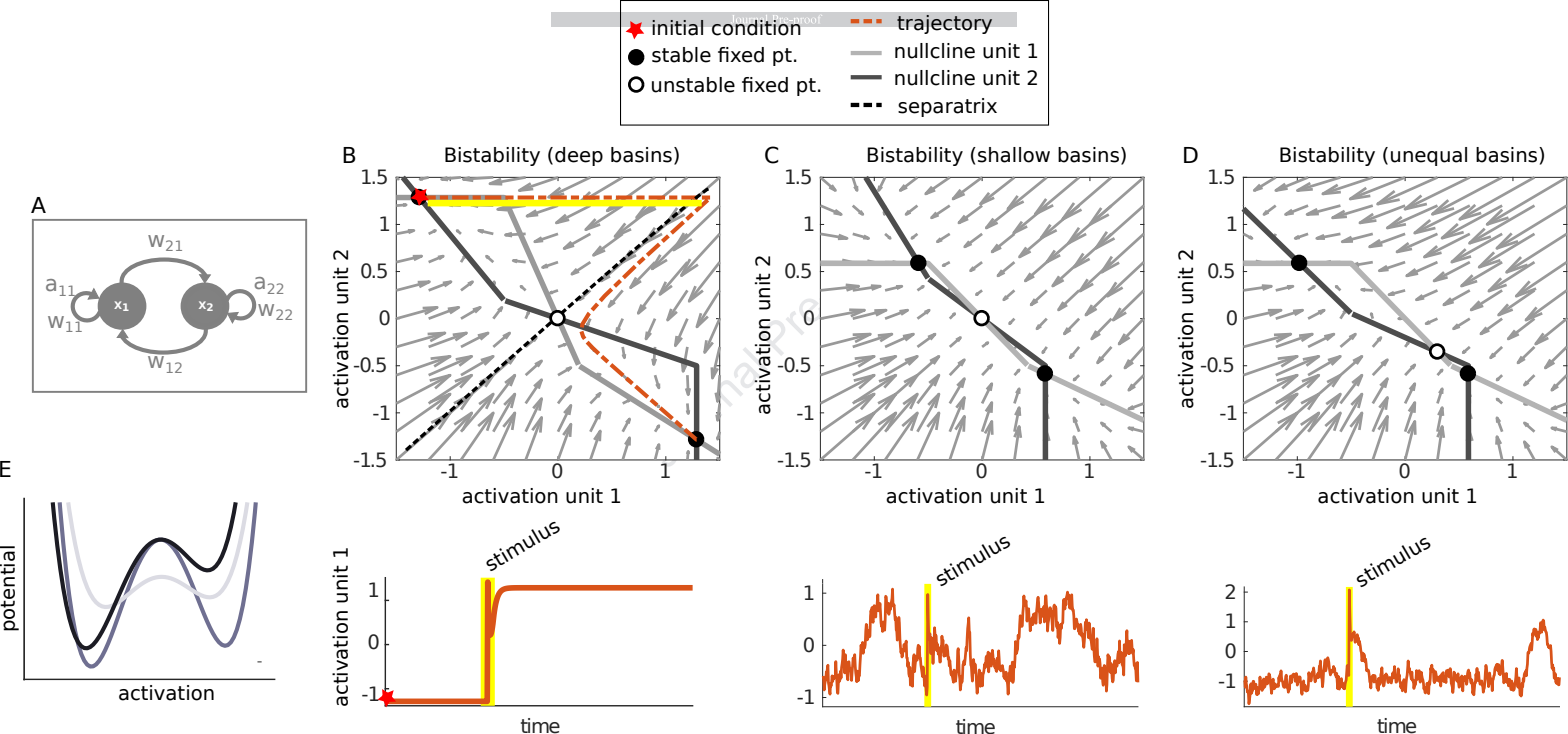
**Fig. 4. Examples of different sequential phenomena in dynamical systems, illustrated with an RNN.** Panels depict flow fields (top row) and time graphs of unit 1 (bottom row) for different parameter settings of an RNN (eq. 6 in Supplement). Gray-shaded lines in flow fields mark nullclines of units 1 and 2, red dotted lines show one trajectory starting from an initial condition (red star), light yellow lines mark the brief presentations of positive external stimuli. A) Bistability among a stable limit cycle surrounding the right unstable fixed point, and a stable fixed point in a 2-unit RNN (Fig. 3A). A brief stimulus to unit 1 takes the system from its stable fixed point to the stable limit cycle. Vectors are all normalized to same length for better visualization of flow direction. B) Heteroclinic orbit (shown in black) connecting the system's two saddle nodes. In this case, the heteroclinic channel (HC) created by this orbit is itself not attracting (in contrast to examples in (69)), such that nearby trajectories tend to diverge from it (but note that, in the deterministic case, the system would continue to move on the heteroclinic orbit if placed exactly on it). Yet, this unstable HC still influences the behavior of the system in the sense that brief perturbations through an external stimulus (shown in light yellow) tend to cause trajectories to move slowly in its vicinity (just below it) until they return to the stable fixed point at the bottom. C) The famous chaotic Lorenz attractor (93), reproduced by an RNN (eq. 6) statistically inferred from trajectories drawn from the Lorenz system (140). As characteristic of chaotic systems, two very close-by initial conditions may lead into very different activation patterns in the longer run, as displayed in the bottom graph for one of the RNN variables.

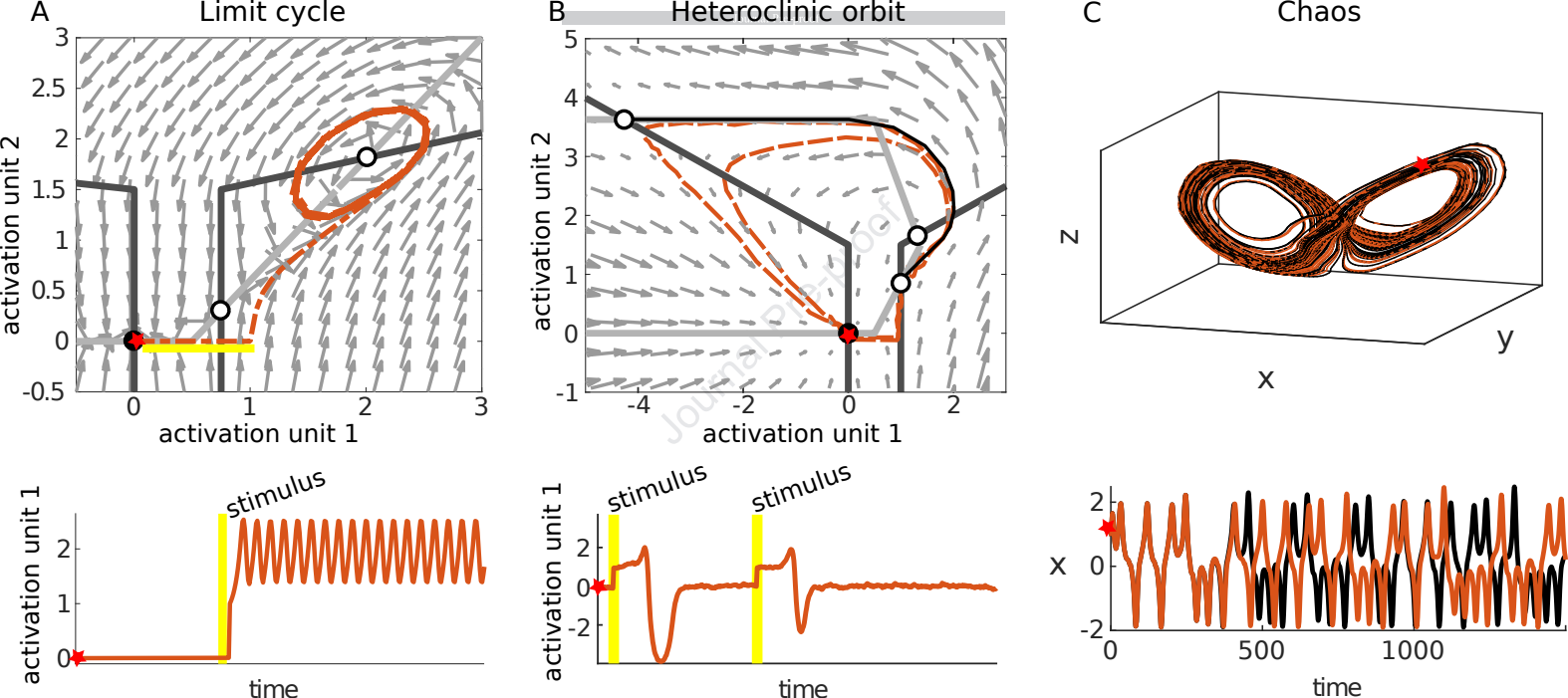
**Fig. 5. Example of line attractors, slow flow, and bifurcations in an RNN.** A), B), and C) depict flow fields for slightly different parameter settings of a 2-unit RNN (eq. 6 and parameters used in Supplement). A) shows a line attractor. Gray-shaded lines mark the nullclines of units 1 and 2. Red dotted lines show one trajectory starting from its initial condition (red star) and briefly pushed away from the line attractor by stimulus pulses (indicated in light yellow), indicating that the line attractor integrates stimuli. B) When the parameters of this line attractor configuration are changed, the system's bottom-right fixed point disappears and leaves behind an 'attractor ghost'. In the vicinity of this attractor ghost the flow is very slow or C) relatively slow, depending how far the system's parameters were moved away from the truly attracting configuration. The bottom figures show the activation of unit 1 for systems in A-C, respectively, starting from the initial condition (red star), and stimulated repeatedly as indicated by the yellow lines. Note that for clarity we omitted trajectories and stimuli from B) and C). The stimuli in B) and C) take the state of the system to the spots in state space indicated by the green crosses.

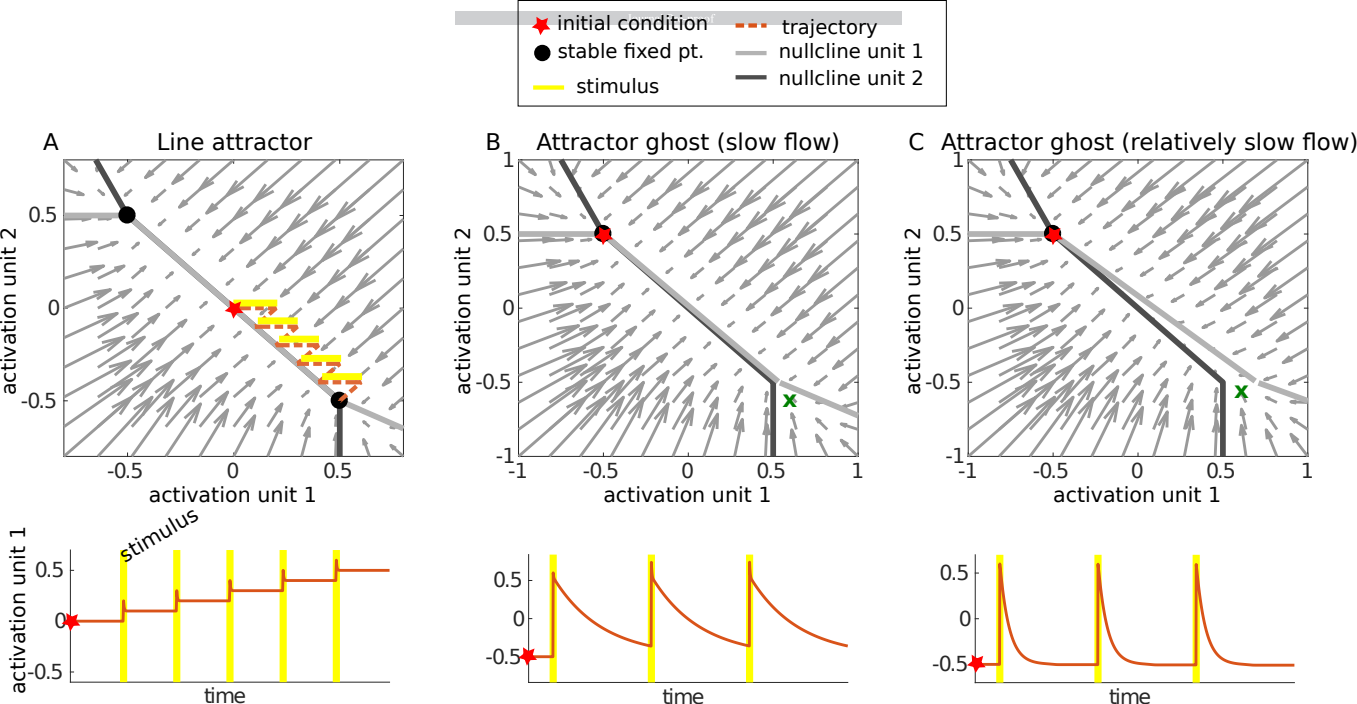
**Fig. 6. Bifurcations.** Example of a bifurcation for the system in Fig. 5A with fixed points plotted as a function of network parameter $\lambda$. Here, $\lambda$ is a factor which regulates the size of the units' self-excitation. With $\lambda<1$, the network exhibits only a single stable fixed point, then switches to a line attractor for $\lambda=1$ (as in 5A), and finally harbors two stable and one unstable fixed points for $\lambda>1$ (as in the systems in Fig. 3).




A
time graphs
state space representation
soc. retreat
stress
mood
time
initial condition
0.5
0
-0.5
-1
mood
0
0.2
0.4
stress
4
2
0
soc. retreat
B
stress
negative affect
time
6
4
2
0
negative affect
0
5
10
stress


initial condition
stable fixed pt.
unstable fixed pt.
trajectory
nullcline unit 1
nullcline unit 2
separatrix
A
$w_{21}$
$a_{11}$
$x_1$
$x_2$
$a_{22}$
$w_{11}$
$w_{22}$
$w_{12}$
B
Bistability (deep basins)
C
Bistability (shallow basins)
D
Bistability (unequal basins)
activation unit 2
activation unit 1
E
potential
activation
stimulus
time


A
Limit cycle
activation unit 2
activation unit 1
B
Heteroclinic orbit
activation unit 2
activation unit 1
C
Chaos
z
y
x
activation unit 1
stimulus
time
activation unit 1
stimulus
stimulus
time
x
time


initial condition
stable fixed pt.
stimulus
trajectory
nullcline unit 1
nullcline unit 2
A
Line attractor
B
Attractor ghost (slow flow)
C
Attractor ghost (relatively slow flow)
activation unit 2
activation unit 1
0.5
0
-0.5
1
-1
stimulus
time