



# How higher goals are constructed and collapse under stress: A hierarchical Bayesian control systems perspective



Rutger Goekoop <sup>a,\*</sup>, Roy de Kleijn <sup>b</sup>

<sup>a</sup> Parnassia Group, PsyQ, Department of Anxiety Disorders, Early Detection and Intervention Team (EDIT), Netherlands

<sup>b</sup> Cognitive Psychology Unit, Leiden University, Netherlands

## ARTICLE INFO

**Keywords:**  
 Organisms  
 Human behavior  
 Hierarchical Bayesian control systems  
 Machine learning  
 Biological networks  
 Network theory  
 Information bottleneck structure  
 Bow-tie motif  
 Variational auto-encoders  
 Hierarchical Bayesian inference  
 Free energy  
 Active inference  
 Entropy  
 Goal-directed learning  
 Goal hierarchy  
 Higher goals  
 Moral functioning  
 Stress  
 Personality  
 Mental disorders  
 Psychopathology

## ABSTRACT

In this paper, we show that organisms can be modeled as hierarchical Bayesian control systems with *small world* and information bottleneck (*bow-tie*) network structure. Such systems combine hierarchical perception with hierarchical goal setting and hierarchical action control. We argue that hierarchical Bayesian control systems produce deep hierarchies of goal states, from which it follows that organisms must have some form of ‘highest goals’. For all organisms, these involve internal (self) models, external (social) models and overarching (normative) models. We show that goal hierarchies tend to decompose in a top-down manner under severe and prolonged levels of stress. This produces behavior that favors short-term and self-referential goals over long term, social and/or normative goals. The collapse of goal hierarchies is universally accompanied by an increase in entropy (disorder) in control systems that can serve as an early warning sign for tipping points (disease or death of the organism). In humans, learning goal hierarchies corresponds to personality development (maturation). The failure of goal hierarchies to mature properly corresponds to personality deficits. A top-down collapse of such hierarchies under stress is identified as a common factor in all forms of episodic mental disorders (psychopathology). The paper concludes by discussing ways of testing these hypotheses empirically.

## 1. Introduction

For centuries, scientists have attempted to discover natural laws that govern the structure and function of living systems. This effort is now producing some interesting results due to theoretical advances, the advent of high-throughput datasets and a huge increase in computing power (Kitano, 2017). Currently, the field still shows a global division between biological and computer sciences, which represents a fundamental distinction in the way the problem has been approached to date, i.e. either by studying living systems themselves (e.g. biology, genetics, biochemistry) or by studying artificial versions of them (e.g. engineering, computer science and robotics). Below, we will first discuss progress

in the fields of artificial systems and biological systems separately. We will then merge insights from both fields to produce a general theory on information processing in living systems and the way they respond to stress. We highlight the universality of this response along with its applicability in humans, and conclude by discussing methods to test the model empirically.

## 2. Artificial systems

### 2.1. Organisms as control systems

Artificial intelligence has now come to a point where computers are

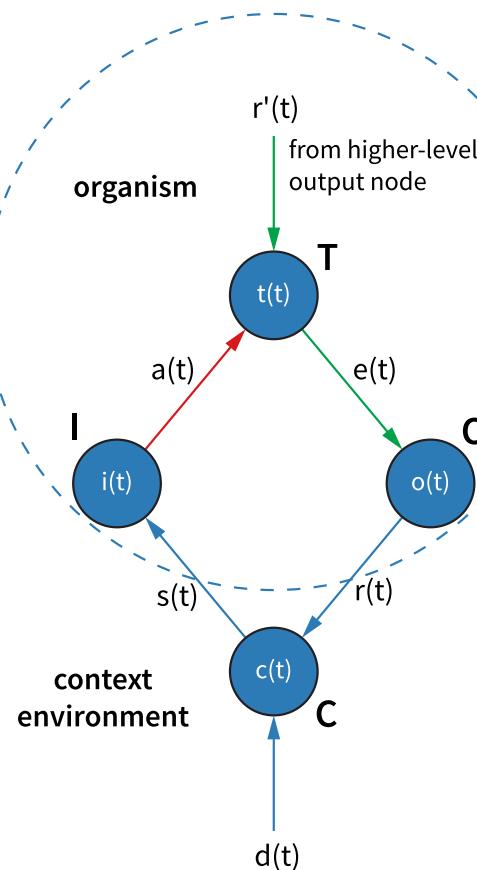
\* Corresponding author at: Parnassia Group, PsyQ, Department of Anxiety Disorders, Early Detection and Intervention Team (EDIT), Lijnbaan 4, 2512VA The Hague, Netherlands.

E-mail address: [R.Goekoop@psyq.nl](mailto:R.Goekoop@psyq.nl) (R. Goekoop).

able to reach (super)human level performance in complex tasks without prior instructions (Mnih et al., 2015; Schmidhuber, 2015; Silver et al., 2017). The basis for this achievement lies in the beginning of the 20th century, when cyberneticists such as W.E. Ashby began to model organisms as control systems (Ashby, 1961; Cannon, 1929, 1932; Powers, 1973a). Such systems maintain internal stability despite changes in environmental conditions by generating some kind of output ( $O$ ) that aims to match the current input state ( $I$ ) with a desired or anticipated throughput state ( $T$ ; a reference value or setpoint). Ashby was the first to highlight the importance of a close coupling between the output and the input of such systems, which is referred to as ‘feedback’. To prove his claims, Ashby constructed a device called a ‘homeostat’, which involved four subsystems that kept each other in check. Each subsystem consisted of a first-order feedback loop that regulated ‘essential variables’ (e.g. blood pressure, glucose levels) and a second-order feedback loop that re-organized a system’s input–output relations when first-order feedback failed, allowing the perturbed system to revert to a stability of its essential variables after all (Seth, 2014). The former, primary form of stability is referred to as homeostasis, whereas the latter form of stability (through additional change) is referred to as ‘allostasis’. This double feedback system is one of the earliest forerunners of ‘hierarchical control’ (see below). By combining four coupled subsystems into one homeostat, the entire control system showed ‘ultrastable’ behavior.

In engineering, control systems are used e.g. in central heating systems, which aim to maintain a stable room temperature despite environmental fluctuations by controlling the radiator. This is done using a control system that compares the current room temperature encoded by a temperature sensor (an input node) to that of a thermostat, which serves as a reference node that encodes a desired temperature (a setpoint). The difference between the two (the error) is transferred in some form to the radiator (an output node), which tries to close the gap between the desired and actual room temperatures (the environment) by emitting heat. Studies indicate that living systems have conditionally independent compartments for input, evaluation and output that allow them to behave in similar ways as control systems (Kirchhoff et al., 2018). Organisms use their senses to monitor the state of their environment and compare their input states to a setpoint state located within a throughput part. The error is then transferred to the output part of the organism, which tries to close the gap between the desired and actual environmental states by generating action (see Fig. 1). Actions then change the state of environment, which feeds back into the senses and the process is repeated. This iterative process helps organisms to find an optimal environmental niche. For example, motor activity in woodlice continues almost ceaselessly and drops to zero only when humidity levels reach near 100 % (a setpoint). As a result, woodlice keep running around erratically until they hit upon a wet place, which is why we find these creatures in all sorts of nooks and crannies. This behavior helps woodlice prevent desiccation and makes them invisible to predators (Friston et al., 2018).

Seminal work by W.T. Powers (1973a,b) showed that biological systems vary their output freely until the state of the input node matches a reference value. Their behavior thus serves to keep a percept (of some environmental condition) within certain limits. Woodlice probably have no clue as to where exactly in the garden they can find a particular crevasse, after which they engage in a carefully controlled output sequence that is aimed at reaching the desired spot. Instead, they just stumble upon a dark and wet place that produces the kind of sensor output that makes motor activity drop to zero. Since Powers considered organisms to control their input (percepts) by means of their output (behavior) and its subsequent effects on the environment, this type of control was called ‘perceptual control’ (Powers, 1973a). Perceptual control theory is highly pragmatic: rather than the specific actions, it’s the end-result that counts. By freely ‘emitting behavior’ (Skinner, 1990) until a desired effect is obtained, organisms can come up with a number of different solutions to the same problem (e.g. running and hiding in crevasses, rolling up, or digging in all prevent desiccation). This adds



**Fig. 1.** Organisms as Control Systems.

Note: Organisms can be modeled as control systems that consist of an input node  $I$  (a sensor), a throughput node  $T$  (a setpoint) and an output node  $O$  (an effector), which are connected by links that symbolize the possibility of energy exchange between these nodes (see text). Arrows show the direction of energy flow, colors indicate positive or negative relationships (red: negative, blue: positive). The sensory node  $I$  has a state  $i(t)$  that is changed as a result of a stimulus  $s(t)$  from the environment  $C$  (context), which is in a changing state  $c(t)$ . The state  $i(t)$  of the sensory node  $I$  is sampled by an afferent connection and the resulting state  $a(t)$  is compared to (i.e. subtracted from) the state  $T(t)$  of a throughput node  $T$  (the setpoint or reference node). The difference (error  $e(t)$ ) between the two states is passed on by efferent connections to the output node  $O$  (in state  $o(t)$ ), which generates the corrective response  $r(t)$  to the environment  $C$ , and so on. External disturbances of the environment  $C$  are modeled by  $d(t)$ . The setpoint of the system  $T$  can be reset by the output from higher level control systems, see text.

flexibility and creativity to the production of behavior (Powers, 1973b). The advantages of perceptual control have been demonstrated in a number of experiments. For instance, robots that run on perceptual control systems can be pushed off their feet in many different ways yet remain stable, whereas robots that run solely on action-control systems can correct their position only in a limited number of ways and tip over (Johnson et al., 2020).

## 2.2. Organisms as hierarchical control systems

Graphical models such as Fig. 1 can produce behavior that can appear quite life-like (Braitenberg, 1984; Powers, 1973a). Nevertheless, such models require an extension in order to explain more complex forms of behavior, i.e. the formation of action sequences that allow organisms to accomplish more complex tasks. For instance, making a cup of coffee involves a number of simple subtasks (‘action primitives’) that need to be placed in a particular order in order to succeed (e.g. heating water, grabbing a cup, pouring hot water over churned coffee beans,

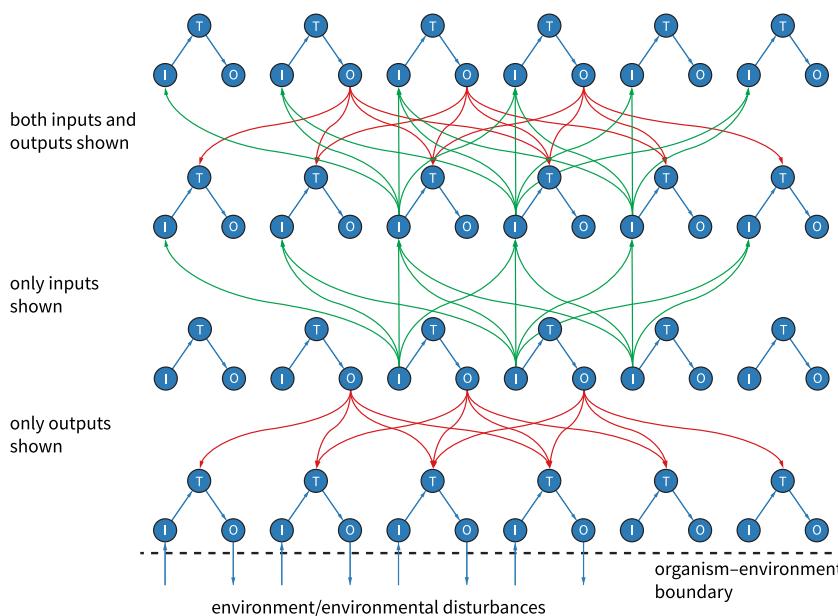
pouring the coffee into a cup, adding milk or sugar, etcetera; Botvinick, 2007). Such output sequences can be more or less efficient depending on the order and the number of recursions in which the subtasks appear (Commons and Pekker, 2008; Solway et al., 2014). Powers showed that perceptual control systems can produce action sequences (behavior) by allowing their setpoints to be reset by the output of other (higher level) control systems and so on, yielding a hierarchy of control systems (Powers, 1973a,b; Powers et al., 1960). This hierarchy is symbolized in Fig. 1 by the input  $r'(t)$  to the setpoint node T. This is the output of a higher-level control system that feeds into the reference signal of a subordinate control system. Fig. 2 shows a more elaborate example of a hierarchical perceptual control system.

The idea that higher order control systems continuously update the setpoints of lower-level systems (to eventually affect the activity of action primitives) is known as the *equilibrium setpoint hypothesis* for motor control (Feldman and Levin, 2009). In hierarchical action control, primitives can be compared to individual musical notes that are activated in parallel ('chords') or in different sequences in order to produce the 'music' of behavior. Studies of hierarchical action control show that action sequences do not require a strict correspondence with the hierarchical wiring of the control system (i.e. we do not engage in a fully hierarchically controlled sequence of coffee-making actions that is spat out from the very beginning of grabbing coffee beans to sipping from the cup; de Kleijn et al., 2014d). Rather, organisms produce intermittent bursts of hierarchically organized action sequences that are updated by a repeated sampling of the environment (action-perception cycles) (Botvinick, 2007). This is comparable with a musician looking up and down at the scroll sheet every now and then to keep track of the piece. Although hierarchical *action* control seems to contradict the notion of hierarchical *perceptual* control, it remains consistent with this notion in the sense that organisms use their (hierarchically controlled) action sequences to eventually control their input states via the environment. Hierarchical action control is routinely used in e.g. robotics, allowing robotic systems to show complex forms of behavior (e.g. Brooks, 1986).

In the past few decades, graphical models of control systems have been modified to explain increasingly complex forms of behavior. Much progress came from studies of reinforcement learning (also termed *operant conditioning*), which added the elements of memory and prediction to control systems (Jordan and Mitchell, 2015; Sutton and Barto, 2018). Such systems update their policies (input-output strategies) depending on the expected reward of some action. The expected reward (a

prediction) is encoded by the setpoints of these systems, of which the state represents the reward or value obtained after a previous action (i.e. a memory). These predictive setpoints are continuously reset (updated) as a function of previous outcomes, keeping track of the values that maximized reward in the past. Thus, reinforcement learning systems iteratively learn the policies that maximize long-term cumulative reward. Whereas earlier systems made no detailed models of the environments they live in (so called *model-free* systems), later systems were allowed to make explicit predictions of the way in which certain imaginary actions would change the input to the system, considering previous experiences (*model-based* systems) (Doll et al., 2012; Solway and Botvinick, 2012). Such 'world models' are simulations of actions and their possible outcomes (e.g. where different paths in a maze lead to and how rewarding that would be), which are based on memories of previous actions and their outcomes. Predictive activity of this type has been compared to the act of planning, imagination, or goal setting, which is why model-based systems are alternatively referred to as *goal-directed* systems. Goal-directed systems require an elaboration of their throughput parts, to accommodate hierarchies of setpoints that encode complex predictive models of the world. Such 'goal states' are continuously updated and pursued by hierarchically organized action sequences until a maximum value has been reached. Studies show that hierarchical model-based systems such as these outperform hierarchical model-free systems in spatial navigation tasks (Botvinick and Weinstein, 2014). This is because such systems construct hierarchies of goals and corresponding subgoals (so called 'goal hierarchies'), which are each pursued in a logical order until the global goal has been reached (e.g. 'get to fruit' = climb tree, jump to other tree, sling to branch, grab the fruit, eat the fruit).

In the past decade, goal-directed learning has been applied within the context of artificial neural networks (Schmidhuber, 2015). Such networks consist of a layer of input nodes that connect to a layer of throughput nodes (a *hidden layer*), which in turn connects to a layer of output nodes. When such systems are trained, the connections within the network are altered until a given input produces a suitable output. It turns out that the performance of such systems increases significantly when their throughput parts are extended to include multiple, hierarchically ordered layers of nodes. Such deep networks can associate raw perceptual input (say, the image of a cat) to a suitable output (e.g. a hierarchical output sequence 'C – A – T') with remarkable precision. When deep networks are allowed to construct explicit world models



**Fig. 2. An Example of a Hierarchical Perceptual Control System.**

*Note:* Classical example of an artificial hierarchical control system, which involves the stacking of one control system on top of another, to produce multiple levels of control. This can be compared to the stacking of one array of thermostats on top of another in order to better control temperature fluctuations in the environment. The output of higher-level control systems can modify the setpoints of subordinate systems (and so on) to produce ordered sequences of action primitives, which we call behavior.

(goal hierarchies), their performance increases even further. Such hierarchical ‘deep belief’ systems can ‘imagine’ a future and formulate efficient sequences of goals and corresponding subgoals that are pursued by means of complex action sequences until the input to the system matches the global goal. Such systems can achieve high success rates (Nagabandi et al., 2018; Pascanu et al., 2017; Racanière et al., 2017; Yamins and DiCarlo, 2016). The performance of these systems comes close to what neuroscientists believe is the essential nature of the human brain: an active inference engine, whose primary job it is to construct predictive models of what is going on in the environment and to test these models by performing some kind of action out into the environment. Such actions change the input to the system (via the environment), which serves as a check on model evidence (Friston, 2010). According to active inference theory, organisms cannot only reduce prediction errors by varying motor output impacting on percepts (as perceptual control would have it), but also by updating their world models to produce a better fit with their input states (a process called ‘Bayesian belief updating’). See below for further information on active inference.

In summary, adding hierarchy to the output parts of control systems allows for the production of complex action-perception sequences (behavioral hierarchies), whereas adding hierarchy to the throughput (goal) parts further boosts the performance of such systems by producing efficient strategies (goal hierarchies). More recently, studies began to apply hierarchical structure to the input layers of deep networks (Mnih et al., 2015; Simonyan and Zisserman, 2014). The hierarchical structure of perceptive areas has been relatively ignored in previous studies, despite the fact that this is a well-known attribute of the cerebral cortex in higher mammals (e.g. receptive fields in the macaque visual cortex) (Hegdé and Felleman, 2007; Rohe and Noppeney, 2015). Hierarchical perception allows control systems to extract increasingly abstract patterns and shapes from raw perceptual input (Karklin and Lewicki, 2009; Kriegeskorte, 2015; Tenenbaum et al., 2011). In 2015, a seminal study was the first to combine hierarchical input (abstract vision) with hierarchical throughput (abstract goal-setting) and hierarchical output (complex action, behavior) to produce human-level performance in complex visuospatial tasks (playing Atari computer games; Mnih et al., 2015). The system only took raw pixel intensity values as input, after which it autonomously discovered complex series of strategies (goals and corresponding subgoals, e.g. taking elaborate detours through a maze) and action sequences (series of jumps and other complex movements) to maximize the outcome of the game (increasing the total score). Similar systems have since shocked the world by beating human experts in activities as diverse as media classification (Simonyan and Zisserman, 2014; Tran et al., 2015), medical diagnostics (Litjens et al., 2017) and the game of Go (Silver et al., 2017) and are quickly finding their way into robotics (Sünderhauf et al., 2018). In short, recent history shows that adding hierarchical structure to the various components of a control system has contributed much to their enormous success.

As illustrated above, the idea that living systems behave as hierarchical control systems is hardly new. Despite its firm rooting within the field of psychology and neuroscience, however, the concept of hierarchical control has been studied largely from the perspective of engineering and computer science, devoting little attention to the finer details of the architecture and function of living systems. Conversely, the idea that biological networks can be modeled as hierarchical control systems has escaped systematic attention in the biological sciences. In the past two decades, there has been a tremendous increase in our knowledge of the structure and function of living systems. This has shown that organisms follow generic rules of structure and function that apply universally to all living systems (see below). These insights have only partly been integrated with the field of control theory and machine learning. The purpose of the current paper is to bring these two influential fields of science further together. We will show that biological systems have a generic network structure that makes them ideally suited

to function as hierarchical Bayesian control systems. Such systems can extract increasing amounts of contextual information from their inner and outer environments, construct increasingly articulated goal hierarchies and generate increasingly complex action sequences in order to reach (long-term) stability. We then identify a universal (stress)response of organisms to contextual cues that overtax their regulatory capacity and ability to remain stable. Such rules can be used to model organisms of any type, including humans.

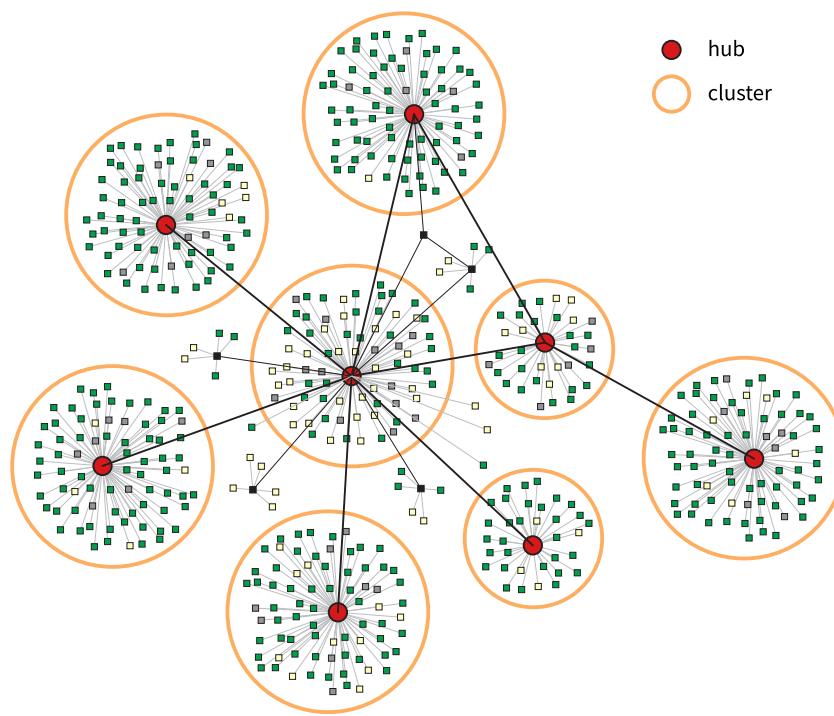
### 3. Biological systems

Network science is booming, ever since the (re)discovery some 20 years ago of the *small world* network structure (Milgram, 1967; Watts and Strogatz, 1998) and the subsequent demonstration that universal laws of network theory govern network structures across a wide range of biological, psychological and social systems (Barabasi, 2013; Barabasi and Bonabeau, 2003; Barabasi and Oltvai, 2004; Barzel and Barabasi, 2013a,b; Newman et al., 2006; Oltvai and Barabasi, 2002). Because of its ability to connect different fields of science using a single methodology and corresponding terminology, network science holds considerable promise as a unifying discipline for many different fields, including biology, ethology, psychology and sociology. Below, we will first summarize some of the main findings from translational network science and identify generic rules of network architecture and function that apply to all living systems. We will then examine generic changes in biological systems when put under severe levels of stress.

#### 3.1. On the structure of biological systems

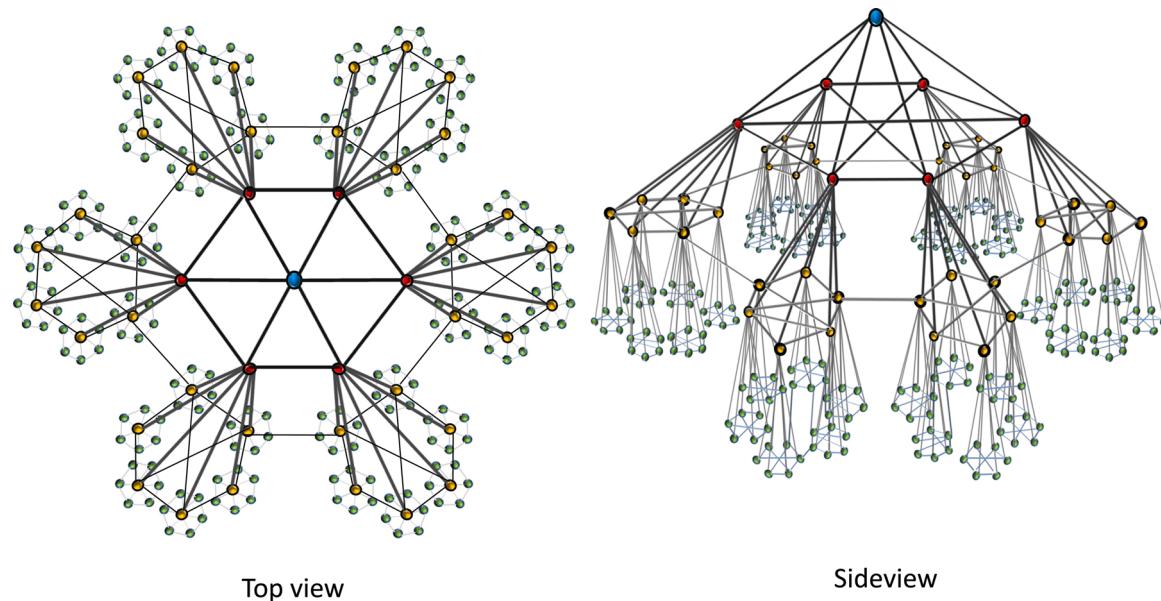
At some level of abstraction, the whole of living nature can be considered to represent the interaction between building blocks that cluster together to form new building blocks, and so forth, until complex multicellular life develops (Oltvai and Barabasi, 2002). Collections of molecules form organelles, which in turn form cells, which in turn form tissues, which in turn form organs, organisms, organizations, biotopes, and so on. At each scale level of biological organization, the interaction between the building blocks that exist at this level (be they organelles, cells, organs, or organisms) can be visualized as a network structure in which building blocks are represented by nodes and their mutual connections by links. Almost without exception, biological networks show a topological structure called the *small world* structure, meaning that most nodes have few connections but some have many (the so-called *hubs*; Fig. 3). Hubs interconnect the various nodes of the network, allowing any two nodes in the network to be connected through a small number of intermediate steps, hence the term ‘*small world*’ (e.g. all people in this world are an average of only 6 degrees of separation apart). Hubs contract large numbers of nodes into densely connected clusters (also called *communities* or *modules*; Girvan and Newman, 2002; Newman, 2006). The nodes that lie within such clusters share more connections amongst themselves than with other nodes within the network, forming subnetworks of their own. *Small world* structures are a general hallmark of biological systems and can be observed throughout living nature (Fig. 3).

*Small world* network structures turn out to be scalable, meaning that network clusters may themselves serve as nodes in a new network structure at a higher scale level of spatial organization, and so on. Thus, biological networks form hierarchies of part-whole relationships, in which higher levels of organization cannot exist without their constituent lower levels of organization (i.e. they form conditional dependencies in space) (Ravasz and Barabasi, 2003). Each new scale level again conforms to a *small world* network structure with multimodular features, which is why this architectural principle is called *scale-invariant*, or *scale-free* (Fig. 4; Barabasi, 2009). The scale-invariance of *small world* network structures has been compared to mathematical constructs called *fractals*: self-similar shapes that follow relatively simple algebraic rules across multiple scale levels of aggregation (Gallos et al., 2007;



**Fig. 3.** Organisms as Small world Network Structures.

**Note:** Organisms can be conceived of as *small world* network structures. In such networks, hub nodes interconnect all other nodes in the network in such a way that the network as a whole has a small average pathlength (i.e. each node is only a small number of steps away from any other node in the network). In *small world networks*, hubs contract parts of the network into communities (modules), which are collections of nodes that share more connections amongst themselves than with other nodes. Because of such features, *small world* networks allow for highly efficient forms of information transfer at low wiring costs with a high tolerance for random damage. They are found in any 'connectome' studied thus far, including genomes, proteomes, metabolomes, microbiomes, neural connectomes, food webs and social networks.



**Fig. 4.** Scale Invariant Structure in Small world Networks.

**Note:** Schematic representation of scale invariant structure in *small world* networks. In such networks, hub nodes contract sets of other (hub) nodes into network clusters. Such clusters may themselves be considered nodes that cluster into superclusters and so on, producing a hierarchy of part-whole relationships. A *small world* network topology (see text) is found at each spatial scale level of biological organization, which is why this topological feature is called ‘scale invariant’ or ‘scale free’. Blue node: central hub, which connects a set of 6 red nodes into a single network cluster. Red nodes are themselves hubs that each contract a set of 6 yellow nodes into another network cluster, etcetera. Note that this process of nested clustering can be repeated almost endlessly, illustrating the concept of scale invariance of *small world* network topology (i.e. any node in his figure may be a network cluster, supercluster, and so on; blue nodes may be drawn into clusters by high level hub nodes, or green nodes may become hubs by adding nodes). Right picture shows a sideview of the left image in which the vertical position of a node indicates its position within a nested hierarchy of hub nodes (a ‘rich club’; [Opsahl et al., 2008](#)).

Song et al., 2005, 2006). Nested modular network structures such as these form spontaneously under the right conditions (i.e. a constant flux of energy into open dissipative systems), since such topologies allow network systems to get rid of their excess energy in the most efficient way, by minimizing resistance to energy flow (Jarman et al., 2017). A basic thermodynamic rule therefore suffices to produce network

structures with short and efficient paths: a phenomenon called *self-organization* (Ashby, 1947; Kauffman, 1993). It has been hypothesized that life kick-started from *small world* networks of chemical reactions, which subsequently adapted to meet the more complex demands of life (Kauffman, 1996; Ramstead et al., 2018).

### 3.2. On the function of biological systems

Biological networks are not just static structures. Energy<sup>1</sup> flows through such structures in the form of electrons, e.g. chemical reactions at the level of receptors and genes, or electromagnetic changes at the level of neurons. In *small world* networks, some parts of the network receive energy (input) from the environment and change their states accordingly. These states are then altered as they flow on through the network in ways that depend on the wiring patterns of the nodes and modules in that part of the network (throughput). The processed states are then passed on to other nodes and modules (output) that lead to some action out into the environment. This succession of state changes is often referred to as network ‘function’. Apart from universal rules of network structure, studies are now beginning to identify rules of network function that apply across different species and scale levels of biological organization (Barzel and Barabasi, 2013a,b; Friston, 2012; Gosak et al., 2018; Kitano, 2004). For instance, the input-throughput-output (I/T/O) organization of most biological networks turns out to resemble the shape of an hourglass, or ‘bow-tie’ (Fig. 5, left image; Csete and Doyle, 2004; Kitano, 2004). The input parts of these structures involve multiple input streams converging onto hub structures, which in turn converge onto higher level hub structures, etcetera, following a hierarchy of part–whole relationships in an upward manner. This goes on until a limited number of high-level hub structures is reached (i.e. the throughput parts). The output parts then involve multiple outputs diverging from these throughput hubs onto lower level hub structures and so on, down the nested hierarchy to the level of individual nodes (Fig. 5, right image). For example, a large number of sensory receptors and corresponding second messenger pathways fan in to a relatively small number of nuclear genes (the waist of the hour glass, or the knot of the bow-tie). Multiple outputs then fan out from these genes in the form of messenger RNAs that instruct a large number of ribosomes to produce all kinds of proteins that are cleaved into even more proteins (Barabasi and Oltvai, 2004; Watson et al., 2015; Zhao et al., 2006). A similar organization can be observed in the human brain (Markov et al., 2013). Here, a large number of neural columns within the visual cortex (coding for color, texture, speed, orientation, etcetera) converge onto a smaller number of brain areas involved in object representations, which in turn converge onto a few brain areas coding for global visuospatial scenes. This convergence goes on until anterior and frontal areas are reached that harbor some of the most global (‘domain general’) representations of the inner and outer environment (the waist of the hourglass). These global states then bias activity levels in several subordinate brain areas involved in the planning and execution of motor programs, which control a multitude of pyramidal cells and muscle fibers to produce motor action (Badcock et al., 2019; Bullmore and

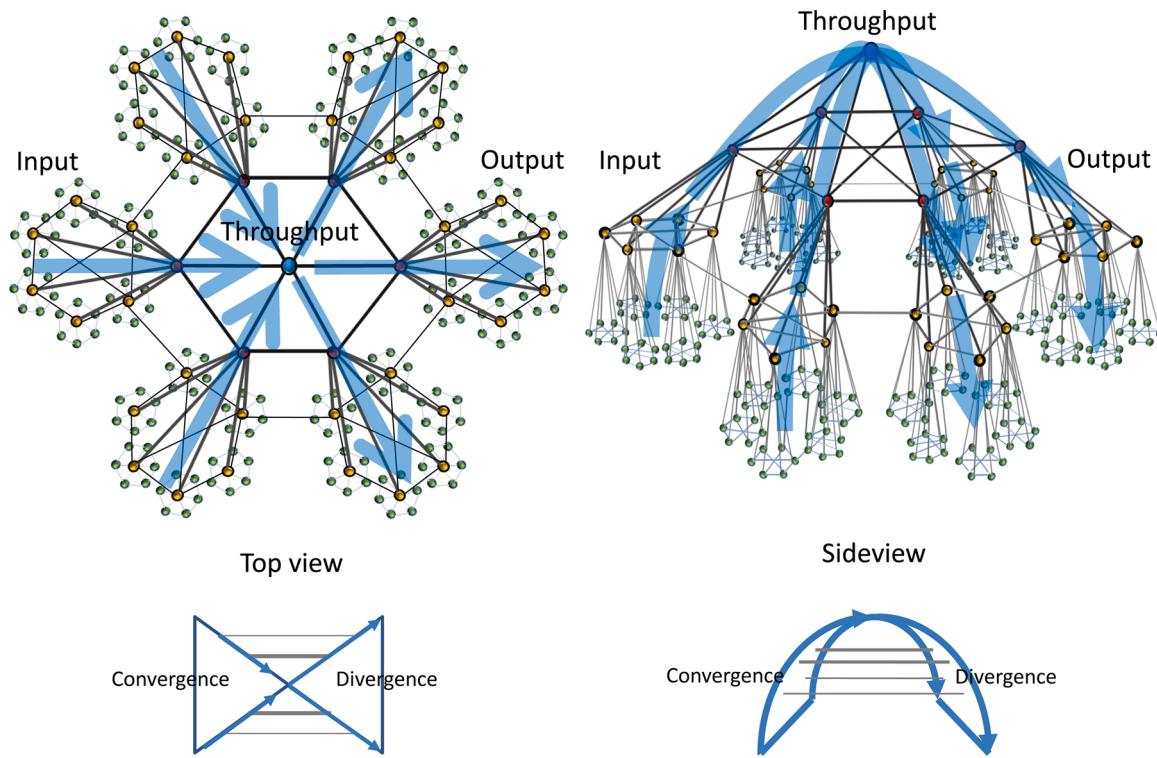
Sporns, 2009; Freeman, 2005; Mesulam, 1998, 2008; Meunier et al., 2010). Bow-tie structures have been observed in the immune system, the internet and within other bow-ties (i.e. bow-ties nested within bow-ties), making this motif a scale invariant phenomenon (Box 1; Friedlander et al., 2015; Kitano, 2004; Zhao et al., 2006).

The ubiquity of the bow-tie motif has sparked questions regarding its functional significance. Bow-ties allow biological networks to convert a host of different inputs into a multitude of outputs using a minimal set of basic operations. Novel inputs and outputs can be easily plugged into a generic core of hub processes without affecting the system as a whole, making it a highly versatile structure. Thus, biological networks can combine robustness with adaptability in a chaotic world full of stimuli (Kitano, 2004). Simulation studies show that hierarchical networks spontaneously evolve bow-tie structure under some restrictions (Friedlander et al., 2015). Resources need to be scarce, and the evolutionary ‘goal’ that these networks aim to satisfy needs to be ‘compressible’, i.e. it should be possible to represent subordinate goal states by an increasingly small number of higher-level variables without losing too much information. This continues until the top of the hierarchy is reached (the knot of the bow-tie, or the waist of the hourglass). The minimal width of the bow-tie structure therefore represents the maximum level of compression of an evolutionary goal, with subordinate structures representing lesser compressed versions of the goal state (Friedlander et al., 2015). As we shall see below, this aspect of bow-ties structures turns out to be rather fundamental: a high-dimensional input is forced through a bottleneck, or low-dimensional manifold. This relates to the concept of dimensionality reduction which can be found throughout statistics and machine learning (e.g. principal component analysis and other clustering methods (Sorzano et al., 2014)). Studies have shown that imposing an ‘information bottleneck’ structure onto hierarchical (deep) networks significantly increases their performance by allowing for some form of compression and generalization of events that take place at lower levels (Hafez-Kolahi and Kasaei, 2019; Shwartz-Ziv and Tishby, 2017). Apparently, living systems minimize complexity cost and use the fewest degrees of freedom to model their environments, i.e. Occam’s principle (Maisto et al., 2015). Organisms can therefore be conceived of as dimension reduction machines that perform a hierarchical clustering on input in an attempt to find the most parsimonious (global) representation without losing too much information. Such high-level compressed representations then fan out to the lower parts of the output hierarchy to produce coordinated action sequences. In short, the bow tie motif provides organisms with an optimal infrastructure to function as hierarchically organized (and model based) control systems.

The flow of information across bow-tie network structures is not a simple process with energy flowing directly from input (via throughput) to output areas in a linear fashion (Kitano, 2004). Bow-tie structures may show cross-connections (shortcuts) between their input and output parts at different levels of the hierarchy, causing the structure to fold back onto itself (Fig. 5, right image). This produces short input-throughput-output loops near the bottom of the hierarchy as well as longer loops that run from input to output along progressively longer throughput loops, reflecting different degrees of processing (Fig. 6). Additionally, feedforward and feedback loops run down and up the hierarchy respectively, reflecting predictive coding as well as corrections of such predictions by means of novel input (Box 1, Fig. 6). Such structures differ from hierarchical control systems that are traditionally used in engineering and machine learning and come with specific functionality. In recent years, insights have grown that organisms are not merely reactive agents that respond passively to external stimuli. Rather, they seem to actively model the causal structure of their inner and outer worlds and use memories to predict future events in a biological equivalent of Bayesian inference (Box 1). The idea that biological organisms engage actively in some form of hierarchical Bayesian inference has produced an explosion of literature in the past decade (Box 2). In this view, each level within a hierarchy generates a predictive model of the hidden causes of the effects (events, activity) observed at a lower

<sup>1</sup> We will use the notion of energy to stand in for the dynamics that couple different nodes or clusters in network graphs. Technically, the energy can be thought of as a log probability of a given state of a node or cluster (i.e. the rarity of a given state, which serves as measure of information content) and the dynamics of network systems can usually be framed in terms of gradient flows on this log probability (i.e. gradient flows on rarity, or information content). A nice example is given by the free energy principle (see below), which defines prediction error as the gradient of variational free energy. In other words, when we talk about energy flows (and network function) we are actually talking about gradient flows on free energy that usually have an interpretation in terms of information flows. As we will see below, this corresponds to predictive coding and Bayesian belief updating in a variational setting (e.g. in artificial or biological systems).

<sup>2</sup> In this paper we refer to (anatomically) backward or descending connections as (control theoretic) feedforward connections, which convey predictions (‘Bayesian beliefs’). Conversely, we refer to (anatomically) forward or ascending connections as (control theoretic) feedback connections, which perform an update on predictions after measurement (this is called ‘Bayesian belief updating’, see below).



**Fig. 5.** Organisms as Nested Modular Small World Networks: The Bow-tie (Hourglass) Motif.

**Note:** **Left image:** organisms can be conceived of as nested modular *small world* network structures with a distinct input-throughput-output organization: a bow-tie (2D) or ‘hourglass’ (3D) structure. The input parts of such networks involve multiple energy streams converging onto each other while ascending in a hierarchy of part-whole relationships (left part of bow-tie). Conversely, the output parts involve multiple energy streams diverging while descending in the hierarchy (right part of bow-tie). The ‘knot’ of the bow-tie (or the waist of the hourglass) lies in between its input and output parts (i.e. the throughput part). This motif can be observed across multiple scale levels of organization, making it a scale invariant feature. **Right image:** bow-tie motifs may show cross-connections (shortcuts) between their input and output parts at different levels within the hierarchy, causing the structure to fold back onto itself (right figure). This allows energy to travel from input and throughput to output structures across loops of various pathlengths, corresponding to different degrees of information processing (see Fig. 6). Please note that arrows in this figure only show the global direction of energy flow. Feedback and feedforward connections<sup>2</sup> run up and down the various levels of the hierarchy, which are thought to represent prediction errors and predictions relative to lower-level input (Box 1).

hierarchical level of organization. The error between the model and the organism’s sensory states served both to update the model and to inspire action, which changes the environment to alter perception, after which the process reiterates. Such generative models perform well across a limited number of observations and trials and their predictions generalize well beyond the subset of training data, suggesting that a certain amount of ‘creativity’ is involved in hierarchical Bayesian modeling (Blei et al., 2017; Tenenbaum et al., 2011). This creative property of higher level models has been linked to the concept of emergence in complex systems (Griffiths et al., 2010; McClelland et al., 2010). So far, however, it has remained largely unclear how this type of processing is implemented in biological systems. This will be discussed in the next section.

### 3.3. Organisms: nested modular network structures that function as hierarchical Bayesian control systems

In the previous section, we saw that organisms are optimally wired to function as hierarchical (nested modular) control systems that combine hierarchical perception (input) with hierarchical goal setting (throughput) and hierarchical action control (output), to iteratively respond to their environments. Below, we will discuss how energy flows travel through such systems (hierarchical message passing) to support Bayesian inference, turning organisms into hierarchical Bayesian control systems.

Organisms do not simply respond randomly to environmental stimuli. Rather, they must connect input patterns to output patterns in ways that are compatible with life, e.g. when the input is ‘food’ (glucose), a

suitable output would be ‘approach’. When the input is ‘predator’ (smell), a suitable response would be ‘avoid’. Such non-random responses are called ‘adaptive’, since they allow organisms to adapt to changing environmental conditions and survive (Gross and Blasius, 2007). Connecting input patterns to adaptive output patterns (‘policy selection’) can be a daunting task for any organism, however. Most organisms live in a rich context of environmental circumstances, which contains multiple cues that may elicit conflicting responses (e.g. approaching food, but avoiding a predator). Such conflicts must be resolved in order to survive (i.e. responses must be prioritized and put in sequence). This requires organisms to build an integrated rather than segregated representation of their environments (e.g. input (food|predator), instead of input (food), input (predator)). Because of their peculiar structure, nested modular (hierarchical) network structures are optimally suited to produce such integrated representations (van den Heuvel et al., 2012v). The input parts of such structures involve multiple inputs that converge onto fewer hub structures (Fig. 5). Like spiders in a web, such hubs keep in touch with the states of large numbers of functionally segregated nodes and clusters in the network, each of which confers part of the relevant information concerning the inner and outer environment. The state of such hubs thus provides a summary representation of the states of all nodes that connect to it (i.e. a state with a higher level of parsimony and abstraction than its subordinate substates, e.g. input (food|predator)). Such functional integration goes on until the top of the nested hierarchy of network clusters has been reached. At each level within the input hierarchy, integrated input states are compared to integrated reference states at a similar hierarchical level (e.g. throughput (food|predator)), after which the ensuing errors are

**Box 1**

## On the Structure of Organisms: Network Motifs and Predictive Modeling

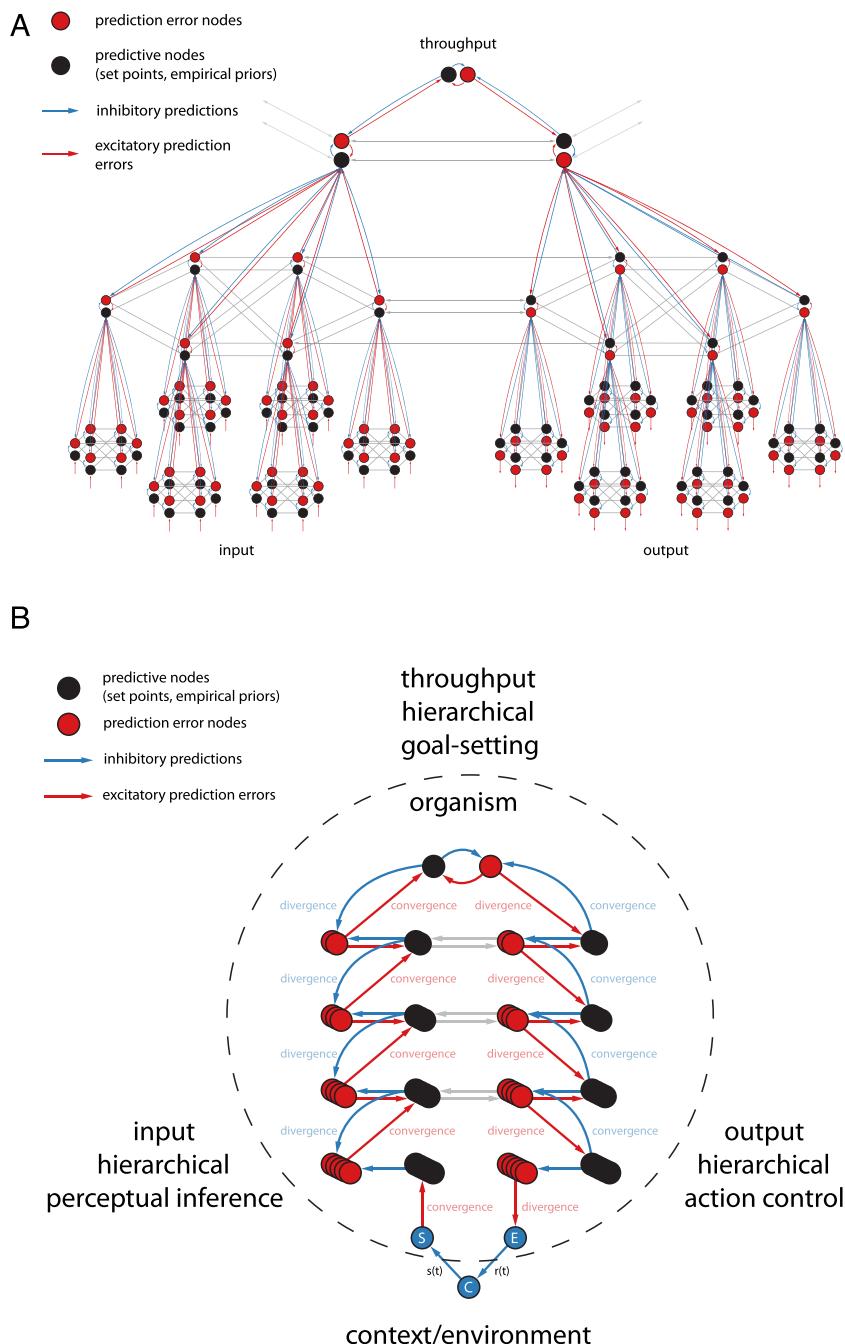
Biological (*small world*) networks are made up of smaller building blocks ('subgraphs') with a relatively large scale called 'network motifs'. These are highly generic pieces of network structure that are observed across different spatial scale levels of biological organization, where they support similar functions (e.g. speeding up or slowing down responses, prolonging responses, integrating or coordinating states, etcetera). The bow-tie structure is just one of these building blocks, with a relatively large size. When examining their finer substructure, bow-ties consist of a family of smaller motifs (Alon, 2007; Araujo & Liotta, 2018; Li et al., 2012). Studies have found a particular abundance of the so called 'feedforward loop' (FFL) in living systems (Alon, 2007). This is a motif that consists of only three nodes (A, B, C) with directed connections between them (i.e. A=>B, B=>C and A=>C). Typically, FFLs lack a connection that runs from the output of the motif back to its input (i.e. C=>A), i.e. they are open loop control systems. When confronted with a stimulus, such motifs push forward a 'best guess' response regardless of its outcome, hence the term 'feedforward'. Because of their ability to forward best guess responses, feedforward motifs have been linked to predictive processing (Del Giudice et al., 2018). For instance, the act of eating already increases insulin secretion regardless of actual increase in blood glucose concentration, which involves a predictive feedforward system (Marchetti et al., 2008; Pezzulo et al., 2015; Pezzulo et al., 2018). In contrast, feedback motifs contain links that run from the output nodes back to the input nodes, i.e. they are closed-loop control systems. Such systems represent events that are the consequences of certain actions. For example, feedback systems are involved in measuring actual blood glucose concentrations after eating, to provide an update on the predictions made by feedforward motifs (Marchetti et al., 2008). The activity of feedforward and feedback systems needs to be balanced in order to have the best of both worlds. In biological systems, FFL motifs represent the feedforward (descending) propagation of predictions from higher levels to lower levels within the nested hierarchy, whereas feedback motifs represent the prediction error that runs back in the opposite (ascending) direction<sup>2</sup>. Thus, feedback and feedforward loops run up and down the bow-tie hierarchy, respectively, to balance prediction errors with predictions. This balance is what underlies 'hierarchical predictive coding' in living systems (see next section, Fig. 6). The ubiquity of FFLs in living systems suggests that predictive activity makes up a substantial part of these projections. This corresponds to cumulative findings that organisms are not merely reactive agents but rather proactive and 'predictive' agents that use memories to predict future events. One of the best known examples is anticipatory salivation in classically conditioned dogs, but Pavlovian learning and anticipatory responses such as these have been demonstrated in organisms as simple as bacteria (Brembs, 2003; Calvo & Friston, 2017; Friston, 2012; Hesp et al., 2019; Mitchell et al., 2009; Tagkopoulos et al., 2008). In short, the nested hierarchical bottleneck structure of bow-tie motifs and their constituent submotifs are a scale free feature, which provides living systems with an optimal infrastructure to function as hierarchical control systems at every scale level of their anatomy (Friston, 2012; Hesp et al., 2019; Ramstead et al., 2018).

conferred to output hubs at a similar level, which then disseminate across lower levels of output hubs, etcetera, to eventually generate complex output sequences (see Section 2).

While encoding their environments, organisms have to solve the 'binding problem' i.e. they need to decide whether signals come from a common cause and should be integrated (i.e. represented by a single node or cluster), or whether they come from independent sources and should be segregated (i.e. represented by separate clusters) (Rohe et al., 2019). The number of independent clusters in a nested modular hierarchy is therefore a function of the number of independent contextual cues that need to be controlled by the organism (Ashby, 1956, Conant and Asby, 1970; Edlund et al., 2011). When environments grow more complex, organisms need to incorporate more clusters in order to produce increasing articulated (contextualized) models. Functional integration across these clusters then increases the hierarchical depth of such systems, allowing for increasingly abstract representations. For instance, some organism can be confronted with food (input A), a mate (input B), a rival (input C) and a predator (input D), all at the same time. It then has to choose whether to eat (output 1), mate (output 2), fight (output 3) or hide (output 4), given its own internal state hungry (input a), alone (input b), wounded (input c), or weak (input d). Each of these factors needs to be encoded into a separate node or cluster ('functional segregation', 'specialization' or 'factorization', e.g. input (A, B, C, D, a, b, c, d)). The functional integration across such perception primitives then produces a hierarchy of part-whole relationships that allows for increasingly abstract (contextualized) percepts when moving up the hierarchy (e.g. input(A|B|C|D|a|b|c|d = input(X)). Similarly, adding primitives to the output parts of a bow tie allows for a richer repertoire of actions (e.g. output(1, 2, 3, 4)). A deeper integration across such clusters produces more elaborate forms of action control and more complex forms of behavior (e.g. 'courtship', which may involve complex action sequences (e.g. output (3|1|4|2) = output (X)). However, extending the repertoire of input-output strategies raises chances that such policies will conflict with one another. In living systems, these conflicts are resolved in a hierarchical fashion (e.g. throughput (A|a) → output(1), throughput (A|B|b) → output(2), throughput (A|B|C|a|b) →

output(3), throughput (A|B|C|D|a|b|c|d) → output(4)). More complex environments therefore require organisms to not only expand their input and output hierarchies, but also their throughput hierarchies, in order to connect input to output strategies in a non-random (adaptive) manner for different combinations of events (i.e. policy selection). In other words, organisms develop hierarchies of reference states and corresponding substates, which are called 'goal hierarchies' (Pezzulo et al., 2015, 2018).

Apart from resolving conflicts between opposing policies in (current) space, goal hierarchies are used to solve potential conflicts in time. For instance, my *currentinput* state  $\text{input}_t(A|B|C)$  (= being warm, well fed, no predators) seems to match my current goal state  $\text{throughput}_t(A|B|C)$  and output pattern  $\text{output}_t(\text{lying down})$ , but this policy may well conflict with my *anticipatedinput* state  $\text{throughput}_{t+1}(D|E|F)$  (e.g. being cold, hungry, lurking predators) and corresponding output  $\text{output}_{t+1}(\text{heating, eating, locomotion})$  (De Kleijn et al., 2014). To resolve such 'temporal' conflicts, the same principle of hierarchical control that allows organisms to integrate increasing numbers of contextual cues in space can be used to integrate contextual cues in time: temporally more distant goal states are encoded by control systems that are superposed onto those that predict temporally more proximal ones in a hierarchy of part-whole relationships (Pezzulo et al., 2018). Errors that are produced relative to such predictive goal states ('prediction errors') may result in actions at a time when such events have not yet taken place (e.g. foraging, stacking fat, storing food, finding shelter, building nests, feeding offspring, preparing to attack). This involves a time and energy investment that is not immediately contingent to the current situation, but serves to keep the system stable through change (i.e. 'allostasis', see introduction). Thus, the ability of organisms to predict events at least some time ahead allows them to engage in 'pre-emptive' actions that significantly raise their chances of survival. The act of anticipating increasingly complex events ever more distantly into the future requires ever deeper hierarchies of goal states, which integrate across multiple levels of subgoals and corresponding timeframes to infer ever more global goal states (Pezzulo et al., 2015). Such highly integrated and predictive goal states are often referred to as 'world models', since they may involve quite complex



**Fig. 6.** Organisms as Hierarchical Bayesian Control Systems with Nested Modular (folded Bow-Tie) Network Structure. Note: Graphical model showing the consensus structure of predictive coding in living systems proposed by (Adams et al., 2013; Friston, 2018; Kanai et al., 2015), which has been adapted to account for the nested modular and folded information bottleneck (bow-tie) network structure that is common to all living systems. **Figure A:** Excerpt of Fig. 5, showing only a single input and output hierarchy for visualization purposes. Each additional level within the nested hierarchy represents a higher level of contextual integration. (Modules of) black nodes encode predictive states (setpoints, updated predictions, goal states). (Modules of) red nodes encode prediction errors. Nested hierarchy of black nodes: predictive (goal) hierarchy. Nested hierarchy of red nodes: hierarchy of empirical evidence. Blue connections: inhibitory predictions. Red connections: excitatory prediction errors. Input hierarchy: predictive hub nodes or clusters suppress (explain away) prediction errors produced at lower levels within the hierarchy through descending and divergent (inhibitory) predictions, reflecting top-down control (e.g. by FFL motifs). The difference (prediction error) is relayed back to higher level predictive nodes or clusters through convergent and excitatory connections, reflecting the bottom-up correction of higher-level predictions (Bayesian belief updating; adjustment of the model, e.g. by feedback motifs). This produces between-level circularly causal dynamics (oscillations). Prediction error and predictive nodes or clusters also engage in circularly causal relationships within the same level of organization, producing within-level oscillations (red and blue arrows, circular shapes). Message passing of the input hierarchy is inverted in the output hierarchy. Here, top-down prediction errors that were not successfully explained away work their way down the hierarchy to supply low-level (predictive) setpoints of action primitives, producing a concerted response. Incidentally, this makes output theory-driven and predictive rather than reactive (Adams et al., 2013). Note that hub nodes (or clusters) of prediction error clusters (or superclusters) within the input hierarchy act as predictive units (empirical priors) at the next level of organization, whereas hub nodes (or clusters) of predictive clusters (or superclusters) within the output hierarchy act as prediction error units. Input and output hierarchies are connected though horizontal connections at different levels within the nested hierarchy (grey connections). This creates longer and shorter loops that run from input via throughput to output, reflecting different degrees of information processing (see text). No horizontal connections exist between the input and output hierarchies at the lowest level of organization, which is an empirical finding (Friston, 2018; Kanai et al., 2015). **Figure B:** Simplified wiring diagram based on connections shown in Figure A, with one more level added when compared to Figure A, adapted from (Kanai et al., 2015). The nested bottleneck (bow-tie) structure is reflected by the copy number of nodes (or clusters), which decreases when ascending in the hierarchy. Horizontal cross-connections (grey) allow energy to travel across loops of different lengths. Short stimulus-response loops correspond to simple (and more complex) reflexes and instinctual responses, whereas progressively longer loops enable habitual and goal-directed behavior. Because of its scale invariance, the entire structure can be seen as one giant (predictive) feed-forward motif. See text for further details. C: context, S: Sensor, e.g. light receptors, E: effector, e.g. striated muscle fiber or mucosal cell, s(t): stimulus, e.g. visual input. r(t): (motor or autonomous) response, e.g. striated muscle action or mucus secretion.

anticipatory designs ('simulations') of the inner and outer environment of the organism, which inspire complex forms of behavior of an increasingly anticipatory nature (see Section 2 and below). Of course, not all organisms equally express goal hierarchies or world models. The height of such hierarchies varies from 'lower' to 'higher' organisms and

between organisms of the same species, causing their behavior to vary along with it.

Rather than generating a predictive model for every possible contingency, organisms use memories to predict future events (e.g. the ringing of a bell causes anticipatory salivation in classically conditioned

**Box 2**

## On the Function of Organisms: Active Inference and the Free Energy Principle

According to the free energy principle, the dynamics of biological systems follows from the basic laws of thermodynamics, i.e. organisms must find their lowest possible energy state despite a continuous influx of energy. In this view, living systems are statistical engines that encode models of the world simply by responding to their input (Friston et al., 2013). The difference between the actual input to the system and some predictive model of the world corresponds to the prediction error of the system, which under some restrictions corresponds to an information theoretic quantity called ‘variational free energy’. Low prediction error corresponds to a low number of alternative states that an organism occupies on average and, therefore, a more stable, low-energy state that has been equated to ‘homeostasis’ (Friston, 2012). Suppressing prediction error is therefore an imperative for all living systems, since it amounts to finding a stable low-energy state. Organisms generally strive towards this overarching goal by generating world models with multiple levels of model complexity and by testing these models against incoming input by performing actions (‘active inference’). Such actions change the environment of the organisms, which produces a novel input that is used as a test on model evidence. In other words, organisms act to maximize sensory evidence for their own predictions: they are ‘self-fulfilling prophecies’. Organisms cannot only reduce prediction errors by changing the environment through action in order to alter their percepts (‘changing your actions’, as in perceptual control), but also by updating their world models to produce a better fit with their input states (‘changing your mind’): a process called ‘Bayesian belief updating’. See text for further details. Although originally formulated within the context of human brain function, the active inference principle has been generalized to involve living systems across multiple spatiotemporal scale levels of organization, varying from microbes and brains to social systems (Ramstead et al., 2018). According to active inference theory, organisms ‘are’ embodied and situationally embedded (Bayesian) models of the world and natural selection is nature’s way of performing Bayesian model selection (Hesp et al., 2019). For equations describing the free energy principle and the process of active inference, see (Friston, 2010, 2012)

dogs). This allows them to restrict predictive modeling to (combinations of) events that have some probability of actually occurring (since they occurred in the past). The act of prediction is therefore intimately tied to the process of learning. Goal hierarchies connect input patterns to output patterns by means of non-random (adaptive) connections. The act of making non-random connections between the input and output of a controls system is called ‘associative learning’. This involves the selective strengthening and weakening of connections within throughput areas, which may e.g. involve molecular bonds in signaling networks or synaptic connections in neural networks. Goal hierarchies develop in the course of an individual’s life, as well as in the course of evolution: any failure to connect stimuli with adaptive responses during the course of their lives (ontogenetic learning) will cause organisms to be eliminated through natural selection (phylogenetic learning). The rewiring of different parts of such hierarchies has been linked to different types of associative learning (Pezzulo et al., 2015, 2018). Short stimulus-response loops represent simple autonomous and/or motor arc reflexes that allow for basic Pavlovian (stimulus–stimulus) learning and instinctive behavior. Longer loops allow for more complex forms of learning such as habit learning and corresponding behavior, whereas the longest loops involve true goal-directed learning and the formation of explicit world models that inspire goal-directed behavior. Goal hierarchies thus consist of progressively longer loops that run from input to output via different levels of integration within the throughput hierarchy. The various forms of associative learning that take place within goal hierarchies are thought to be universal to organisms at any scale level, with more (spatially and temporally) integrated forms of learning occurring within increasingly ‘higher’ organisms. Pavlovian (predictive) learning has been observed to occur within organisms as primitive as bacteria (Calvo and Baluška, 2015), whereas goal-directed learning is observed within higher vertebrates and some invertebrates (Pezzulo, 2012).

Although it is now increasingly recognized that organisms are predictive agents (see Box 1), it remains unclear how exactly predictive modeling is implemented in living systems. Having a nested modular network structure with information bottlenecks motifs appears to be a necessary precondition, but it is not a sufficient one. The graphical model of Fig. 4 therefore requires modification to allow for predictive coding. To this end, tentative hypotheses have been put forward that are based on hierarchical message passing in the human brain (Adams et al., 2013; Friston, 2018, 2019b; Friston et al., 2017; Kanai et al., 2015; Kiebel and Friston, 2009). In Fig. 6, we show the putative wiring scheme for hierarchical predictive coding in biological systems (Adams et al.,

2013; Friston, 2018; Kanai et al., 2015), which we adapted to accommodate a folded information bottleneck structure (a ‘bow-tie’ motif). Here, predictive states are encoded by nodes at a higher level of integration, which suppress prediction errors at lower levels of integration by means of divergent (disynaptic) inhibitory connections (Fig. 6). The difference (prediction error) is conveyed horizontally to the output hierarchy as well as projected back upward by convergent excitatory connections to correct these higher-level predictions (update the models), turning them into posterior expectations (‘empirical priors’). This process is called ‘Bayesian belief updating’ and involves the actual learning process (i.e. a change in connective efficacy). Thus, higher level models attempt to suppress (‘explain away’) prediction errors produced by lower-level systems, whereas lower-level systems in turn correct higher-level predictions. Such circularly causal relationships produce oscillations that are typically observed in neural dynamics. For an overview of the mathematics describing the process of hierarchical message passing in the context of Bayesian inference, see (Kiebel and Friston, 2009).

The output hierarchy shows a similar but inverted makeup in exactly the same way described under the equilibrium setpoint hypothesis, or indeed perceptual control theory (Adams et al., 2013; Friston, 2019b). Here, prediction errors descend down the hierarchy while diverging onto lower-level hub nodes to correct low-level predictive models, whereas predictions ascend up the hierarchy while converging onto higher-level prediction error units. Thus, prediction errors globally ascend and converge within the input hierarchy and descend and diverge within the output hierarchy, to eventually supply the setpoints of lower-level output primitives (e.g. motor or autonomous reflex arcs). Each level within the input hierarchy tries to explain away prediction errors produced at lower levels within the hierarchy by means of inhibitory (predictive) connections (Fig. 6). If prediction errors cannot be suppressed by a simple (less integrated) world model and corresponding output produced at the bottom of the hierarchy, they are carried up to the next level in an attempt to suppress the errors using a more elaborate (contextually more integrated) model (see section 3.5). In action control, this process of hierarchical message passing takes place in inverted order. Here, prediction errors that have not been successfully explained away run down the hierarchy to inspire action. Such output may still reduce prediction errors within the input hierarchy by changing the environment and, hence, the input to the system (‘active inference’).

The process of predictive coding and belief updating as described above is thought to reflect hierarchical Bayesian inference in biological

systems, and can be seen as a general model for information processing. It is thought that similar principles apply in any organism, from microbe to man (Friston, 2012, 2018; Hesp et al., 2019; Ramstead et al., 2018). For instance, membrane receptors and second messenger pathways may represent posterior expectations that are informed by genetic or biochemical priors (setpoints) at different levels to produce output. Such systems may produce oscillatory dynamics similar to those observed in neural dynamics (Friston, 2012). As can be seen in Fig. 6, information bottleneck motifs can be observed at the level of individual nodes, clusters, superclusters and within the network structure at large, i.e. it is a scale invariant feature. As a consequence, the organism itself can be modeled as one giant feedforward loop motif (Box 1), which produces predictive output that feeds back into the organism through the environment, providing an update on predictions ('active inference', Box 2). This means that Fig. 1 should be adapted to contain an arrow running from node A directly to node C.

At this point, it is important to emphasize the difference between traditional notions of hierarchical Bayesian inference in statistics and hierarchical inference as it takes place in living systems. First, statistical models usually involve a single hierarchical generative model. In living systems, the architecture of generative models acquires two streams: a sensory or input stream that controls input while is primarily concerned with inferring "what the world is doing" and an executive or output stream that tries to infer "what the organism is doing" (either in terms of motor behavior or autonomic function): the dual hierarchy in Fig. 6. Input hierarchies are involved in hierarchical perceptual inference, i.e. producing increasingly comprehensive perceptual models that try to explain lower-level sensory events (Friston et al., 2006). Output hierarchies on the other hand are involved in hierarchical action control, i.e. decoding high-level abstract models within the information bottlenecks of organisms into detailed action sequences produced by action primitives located at the base of the output hierarchy. Unexplained (residual) prediction error thus moves down the hierarchy to eventually supply the setpoints of low-level action primitives and pose as complex 'output commands'. Meanwhile, predictions with respect to the hidden causes of sensory events that take place in motor (e.g. proprioceptive) or endocrine (e.g. interoceptive) structures run upward in this hierarchy, in an attempt to suppress prediction errors. This counterstream represents feedback on the correct execution of motor or endocrine actions, based on the organism's models of what it is doing (i.e. based on the inferred sensory states of its output organs). When predictions with respect to the actual state of output organs (represented by bottom-up predictions) matches the output command (by top-down prediction error), prediction errors are fully suppressed and the execution of the output pattern comes to a halt. (Friston, 2019b). This dual aspect of hierarchical inference is emphasized by referring to nested hierarchical bow-tie network architectures (with small-world characteristics). This means that "bow-tie" should be read as a dual-aspect spatial hierarchy responsible for making inferences both about hidden states of the world and actions upon those states, respectively.

Second, models of hierarchical (Bayesian) inference in statistics are unfamiliar with the concept of goal-directedness (agency). This concept is still a topic of debate (Walsh, 2015), yet seems to be clearly definable from the perspective of organisms as hierarchical control systems. As observed in Section 2, perceptual control theory already equated the reference signal (setpoint) of control systems with goal-directedness and the hierarchical organization of reference signals with the formation of more complex goal states (Powers, 1973b). Similarly, model-based control theory involves organisms constructing elaborate hierarchical models of the world that serve as predictive goal-states that are encoded by intermediate throughput areas (Solway and Botvinick, 2012). In active inference theory, goal states align with so called empirical priors. These are nodes or clusters that encode prior beliefs that have been updated by sensory input, i.e. priors at intermediate levels within a hierarchical model (the black nodes and clusters in Fig. 6). Such nodes or clusters encode the states, or sensory information sampled, that the

organism *a priori* prefers to occupy or sample, after having been updated by a certain input (red nodes in Fig. 6). Goal states can therefore be construed as 'posterior expectations and beliefs about controllable but hidden states of the world'. The scale free nature of living systems makes sure that empirical priors form nested hierarchies, with higher-level clusters of priors reflecting increasing amounts of contextual integration of preferential or predictive states (i.e. from individual setpoints to complex world models). In other words, the nested modular hierarchy of black nodes and clusters in Fig. 6 (empirical priors) reflects a hierarchy of goals and corresponding subgoals, down to the level of individual setpoints. Similarly, the nested modular hierarchy of red nodes and clusters in Fig. 6 (prediction error units) represents a hierarchy of empirical evidence at different levels of contextual integration, which is aligned along the various levels of the goal hierarchy to provide an update on these models. Thus, instead of being fixed and given, goal states are progressively inferred within the narrowing bottlenecks of bow-tie structures that form a smooth continuum between input- and output hierarchies (Fig. 6A). These structures are involved in inferring "what the organism should be doing", i.e. hierarchical goal-setting. As a result, we necessarily introduce the notion of 'hierarchical Bayesian control systems'. Such systems combine hierarchical perceptual inference (input) with hierarchical goal inference (throughput) and hierarchical action control (output), to eventually reduce overall levels prediction error through active niche exploration ('active inference').

This concludes our description of how goal hierarchies are constructed in living control systems. Below, we will examine which global types of goal states are produced within deep goal hierarchies and discuss their putative positions within a nested hierarchy of network clusters. We will then show how such hierarchies collapse in a top-down manner under rising levels of stress, leading to corresponding changes in behavior.

### 3.4. A taxonomy of goal states

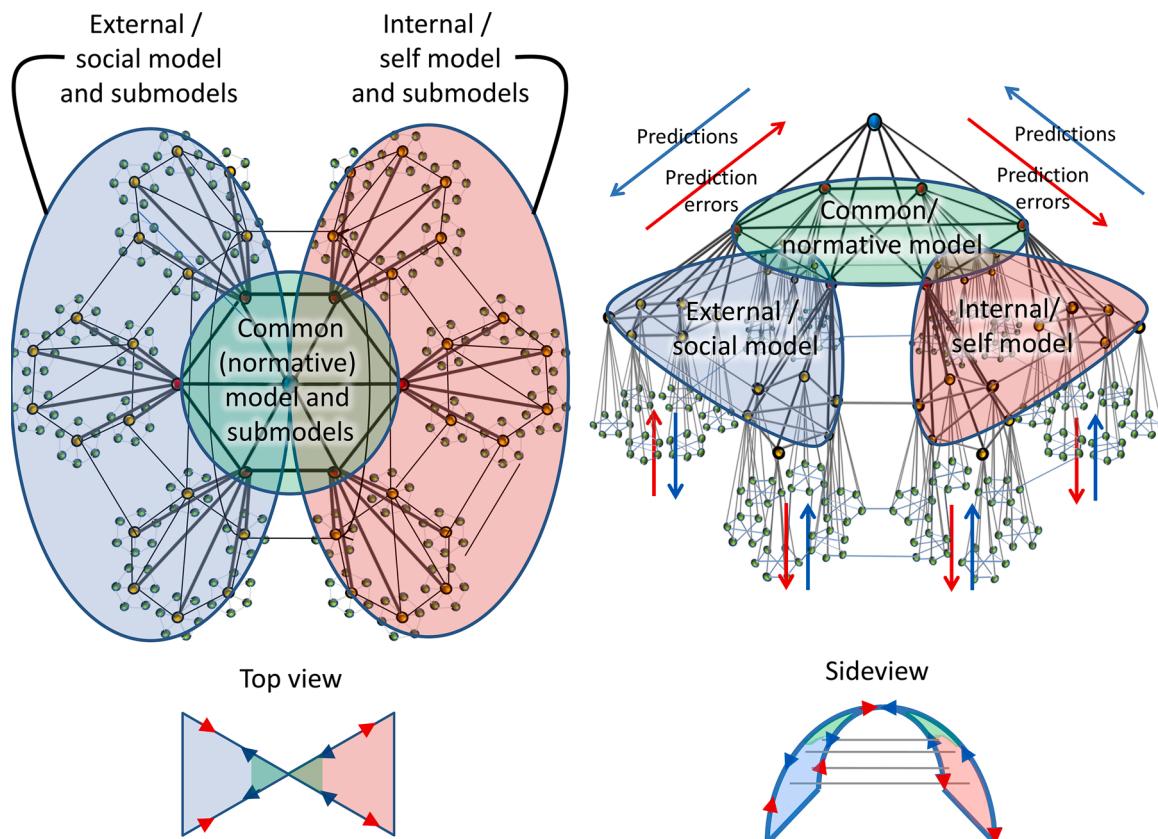
The central tenet of hierarchical Bayesian inference in biological networks is that organisms try to infer the hidden causes of their sensory input (effects) and construct predictive models to do so. The difference (error) between these predictions and the perceived events is used simultaneously to inform behavior (output) and to adjust the model (Friston, 2012). From both observational data and theoretical considerations, organisms are thought to construct at least two global types of predictive models (goal states) at the top of their goal hierarchies. One of these involves a model of the organism itself (Limanowski and Blanckenburg, 2013; Moutoussis et al., 2014). Since any organism has a body, it will consistently receive input that can be explained as produced by or originating from within that body. Such signals may involve both changes in the internal state of the organism (e.g. changes in internal glucose or acidity levels) as well as changes in its external environment as a result of actions produced by the organism itself (e.g. chemicals secreted or vibrations produced by the organism itself). Through hierarchical Bayesian modeling, organisms will eventually infer the hidden common causes behind these various types of signals (effects) and, eventually, the 'self' as a single common cause. Prediction errors relative to such 'self-models' produce behavior that shows hints of a sense of agency (e.g. a differential response to signals produced by the organism itself rather than its environment). The principle of hierarchical Bayesian inference therefore predicts that self-models are produced to varying degrees in any organism, from microbe to man. Most organisms have different sensory systems for monitoring their inner and outer worlds. 'Exterocepis' is used for sensing external events and usually includes 'the 5 senses', i.e. vision, smell, hearing, touch and taste. Interocepis is used to monitor internal events and involves sensory streams from smooth muscles, endocrine glands and other organs. Propriocepis is used to relay the state of the world in between the interior and the exterior of the organism and involves data streams from striated muscles (e.g. muscle spindles). Internal/self-models rely

disproportionally on sensory information derived from internal systems (e.g. interoception and proprioception) (Limanowski and Blankenburg, 2013; Moutoussis et al., 2014). Since output hierarchies are primarily involved in inferring ‘what the organism is doing’ and rely disproportionately on internal sensory streams to do so, internal/self models likely form a continuum with the top of the output hierarchies of bow-tie structures (Figs. 6 and 7). This at least seems to be the case in the human brain (Seth and Friston, 2016; Smith et al., 2019a; Thornton et al., 2019).

Similarly, organisms can infer the (hidden) common causes of effects (input) produced by factors outside of their own body (Baker et al., 2017; Limanowski and Blankenburg, 2013; Ondobaka et al., 2017). Apart from abiotic factors such as rain or snow, such external factors may involve models of other organisms and their intentions (e.g. predator, prey, friend or rival). Such social models are produced to varying degrees in any organism. Prediction errors relative to these models inspire social behavior, which reflects some sense or knowledge of the agency of other organisms, i.e. their existence, social roles, needs and intentions. Such behavior can be found already at the level of bacteria (e.g. quorum sensing in biofilms; Lyon, 2015). External/social models predominantly rely on sensory information derived from input organs (e.g. exteroception) to derive the state of the external world (Smith et al., 2019b; Moutoussis et al., 2014). Such information is then passed on to internal systems to formulate an adaptive response and monitor its execution. Meanwhile, external/social models control the output of the same external systems in a hierarchical manner (i.e. attentional biasing of exteroceptive organs). Since input hierarchies are concerned with

inferring what the external world is doing, it is therefore likely that external/social systems form a continuum with the top of the input hierarchies of bow-tie structures whilst being strongly connected to internal/self systems (Fig. 7). Again, this at least seems to be the case in the human brain (Seth and Friston, 2016; Smith et al., 2019a; Thornton et al., 2019).

As observed in section 3.3, the complexity of a goal hierarchy may vary across individuals and species depending on environmental complexity, and the behavior of their owners varies along with it. We therefore predict that organisms that display a greater degree of agency should show a local extension of their nested hierarchical trees to encode more explicit self-models, i.e. involve the integration across a larger number of network communities. This hypothesis can be tested e.g. by examining organisms that differ in the degree to which they respond differentially to (chemical or physical) signals produced by themselves rather than their environment, or the degree to which they show signs of (self-referential and goal-directed) behavior (agency). Such organisms should have larger scores on measures of hierarchical depth within specific parts of their networks (see Discussion). Similarly, we propose that social behavior, when compared to solitary behavior, should involve some local extension of their hierarchical trees to encode more explicit social models. Such models may become especially intricate in highly sociable species that spend a lot of time gauging the social roles and intentions of their community members (e.g. some birds, mammals and primates). Such organisms are constructing world models of the world models of other organisms (i.e. recursion and reciprocity; Friston and Frith, 2015). These hypotheses can be tested by comparing the



**Fig. 7.** Putative Relative Positions of High-level Goal States Within Living Network Systems.

*Note:* Schematic view of the way in which higher level ('normative') world models may develop within a hierarchy of goal states through the functional integration across self and social models. External (social) models predominantly involve inference on exteroception and may hence form a continuum with the input-part of a goal hierarchy. Internal (self) models predominantly involve inference of interoception and proprioception and may hence form a continuum with the output part of a goal hierarchy. Logically, cross-cutting (normative) models that integrate across internal and external world models and corresponding time domains form the top of the goal-hierarchy (i.e. the highest level of inference). For visualization purposes, no differentiation is made between predictive and prediction error nodes or clusters (for details on this, see Fig. 6). Individual nodes in this figure may represent both single nodes and clusters, conforming to the scale invariant principle.

hierarchical network structure of solitary and social species, or social species that differ in their level of sociability (see Discussion). A similar argument can be made for the ability to predict events ever more distantly into the future. We predict that temporally more distant goal states require deeper hierarchies of control i.e. the integration across a larger number of network communities. This can be seen as a hierarchical extension of interior (self-referential) and/or exterior (social) models to accommodate long-term predictions with respect to self and/or others. Such anticipatory actions may be aimed at a future version of the individual itself or some external agent, rather than the current self or the current other. This hypothesis can be tested by comparing the hierarchical depth of network (bow-tie) structure between individuals or species that differ in their ability to anticipate (self-referential and social) events (see Discussion).

In hierarchical Bayesian inference, each superordinate level performs a form of ‘pattern recognition’ on events that take place at subordinate levels. The superordinate level thus encodes a more generalized and parsimonious (abstract) model of events that happen below it. Such higher-level generative models go well beyond the lower-level data that helped to spark their existence: they may involve quite creative designs that may autonomously inform behavior (Tenenbaum et al., 2011). When this principle is applied systematically to goal states, something interesting happens. As mentioned, organisms produce a hierarchy of goal states that eventually involves a global division between internal (self-referential) and external (social) goal states, both of which can be set proximally or more distally in time. Logically then, the hierarchical integration across goal states can be pushed one level further, involving an additional level of inference across these two global goal states. This produces an overarching third series of goal states that are common to both the organism itself and its (social) environment, across timescales (Fig. 7).

Such models transcend the level of the individual organism, its immediate (social) environment, as well as the immediate moment. In other words, such goal states define (social) laws, rules or standards that hold across different individuals, social groups and timescales (Constant et al., 2019; Toelch and Dolan, 2015). Thus, hierarchical Bayesian inference predicts that, eventually, organisms produce goal states that they consider to have general validity for everyone across (infinite) time. Prediction errors that are produced relative to such ‘normative’ goal states may involve a time and energy investment that is not immediately contingent to the interests of the organism itself. Rather, such behavior is aimed at striking a balance between the short-term and long-term interests of individuals and ever more distant social groups (including future generations), i.e. to promote global rather than local stability. Individuals that follow such goals will at times make decisions that favor the (long-term) interests of others rather than themselves, i.e. they will show altruistic behavior. Additionally, such goals may cause some members of a group to punish themselves or others for social norm violation (Fehr and Schurtenberger, 2018). Altruistic and law-abiding forms of behavior have been observed in a variety of (higher) organisms (e.g. Bekoff and Pierce, 2009). We expect such goal states to represent the highest level of hierarchical Bayesian inference and, therefore, the highest level of integration within a nested hierarchy of network clusters. In other words, they truly represent our ‘highest goals’.

This prediction can be tested by examining organisms that differ in the degree to which they engage in activities that are aimed at promoting global and long-term rather than local and short-term stability of individuals and groups (e.g. mediation versus social polarization, fairness versus unfairness in the sharing of energy and resources, punishment for social norm violation versus laxity, altruistic versus selfish behavior, transpersonal identification versus nepotism, transgenerational identification versus generational individualism, etcetera). Combinations of such functions are typically (but not exclusively) found in so called ‘higher organisms’, and should link to measures of hierarchical depth in nested modular biological networks (see Discussion). Thus,

hierarchical Bayesian inference may explain why higher organisms tend to have bigger throughput areas (e.g. the giant nucleus of eukaryotic versus prokaryotic cells, or the frontal and anterior extensions of the brains of higher primates): such hierarchies are required to accommodate more encompassing world models. Despite such extensions, however, the basic principles that govern behavior in higher organisms appear to be the same as in woodlice: action sequences are produced that aim to minimize prediction error relative to world models with different degrees of model complexity (Botvinick and Weinstein, 2014; Friston, 2012).

In humans, empirical studies of goal states have produced a hierarchical taxonomy that eventually involves the global goals of agency (connecting with the self), communion (to connect with a local social group) and meaning (connecting across spatial, temporal and social barriers; Talevich et al., 2017). These global goals are closely related to Maslow’s hierarchy of needs (with multiple levels of self-actualization, social belonging and transcendence) (Koltko-Rivera, 2006). Such goal states have a strong resemblance to internal (self), external (social) and cross-cutting (normative) goal states as predicted by hierarchical Bayesian inference. By now, the human mental phenotype has been mapped quite well with respect to the presence of normative functions and individual differences in the degree to which subjects score on these phenotypical dimensions can explain differences in normative or moral behavior (e.g. Koltko-Rivera (2006); Stankov (2007); von Collani and Grumm (2009)) as well as individual differences in brain structure and function (see below). The existence of such domains of functioning has been eschewed by scientists for quite some time because of its inherently moral (or even religious) nature. Nevertheless, such domains are predicted by the principle of hierarchical Bayesian inference and supported by evidence from various domains of science.

In the active inference literature, goals are prior beliefs about controllable factors in the environment that rest upon each organism’s place in a particular eco-niche, with each niche showing varying degrees of pro-sociality. This leads to the notion of variational eco-niche construction, whereby each individual builds its own generative models that can be shared among other members of its family or conspecifics (Constant et al., 2018; Veissière et al., 2019). The notion of higher (interpersonal) goals amounts to a shared generative model or narrative that ensures the members of a group can predict each other - and thereby minimize their prediction errors. In humans, the need to exchange such higher-level insights gives rise to our scientific, moral and legal institutions, which may aid in the attempt to eventually construct a globally held world view that serves to optimally inform human behavior. Below, we will discuss how stress causes organisms to downgrade on contextual processing (model complexity) and discuss its impact on (human) behavior.

### 3.5. Stress in hierarchical Bayesian control systems

The hierarchical Bayesian control systems perspective on living organisms allows for a clear definition of ‘stress’ (Peters et al., 2017). Stress can be defined as the difference between some desired or anticipated state (a setpoint, goal state, or world model) and the actual input state of an organism. Mathematically, this can be framed as the differences between empirical priors and posterior expectations, i.e. the overall level of prediction error. Likewise, the stress response can be defined as the behavior that follows these prediction errors (note that according to this definition, every prediction error is a form of stress, and any response can be defined as a stress response). (Stress) responses serve to counter the perturbation of a control system and allow the system to return to a more stable state. Since organisms occupy a wide range of environmental niches in which they face a multitude of unique stressors (e.g. specific chemical constitutions, rivals or predators), stress responses are in many cases unique and involve unique messaging pathways. Some stress responses are more general, however, such as the stringent response in bacteria when subjected to nutrient deprivation

(Boutte and Crosson, 2013), or the SOS response in case of DNA damage (Baharoglu and Mazel, 2014). Such ‘general stress responses’ are mounted in a similar way across species, regardless of the specific stressor the organism encounters (e.g. starvation, drought, heat, cold, acidity, salinity, DNA damage, social stress) and involve the up- or downregulation of a few key transcription factors that have been tightly conserved throughout evolution (de Nadal et al., 2011d; López-Maury et al., 2008; Lyon, 2015; Marles-Wright et al., 2008; Nagar et al., 2016; Storz and Hengge, 2010). Whereas the upstream changes in transcription factors that occur during general stress responses show a clear overlap between species, however, the downstream changes appear to be more species-specific. Additionally, general stress responses have only been characterized in a small number of model organisms, so it is uncertain whether they occur within all species of bacteria, or indeed in other species (Gottesman, 2019). At first glance, then, the sheer heterogeneity of stress responses seems to deny the existence of a truly ‘general stress response’. Despite such heterogeneity, however, recent studies found evidence that some aspects of the stress response are indeed universal across species, whether they be bacteria or plants, mammals or humans. This condition- and species agnostic response can be quantified in terms of the overall amount of regulatory activity that takes place under stressful conditions. Studies show that bacteria that are challenged with an evolutionary familiar stressor show subtle responses of gene transcription, whereas bacteria that are challenged with a relatively unfamiliar stressor (e.g. an antibiotic) show larger and seemingly more chaotic responses (Jensen et al., 2017). More specifically, stressed bacteria express a larger number of different genes with increasing amplitudes, while gene expression is becoming increasingly uncoordinated as the challenge endures. These changes have recently been quantified in terms of entropy, which is a well-established information theoretic quantity of disorder (i.e.  $H = -\ln(|M|)$  where  $M$  is the (permuted) covariance matrix containing gene co-expression strengths) (Zhu et al., 2020). Rising entropy levels in the signaling pathways of bacteria successfully predict bacterial fitness in terms of growth rate (stagnation) and survival (death) under stressful conditions. This is true regardless of the specific environmental conditions, the types of genes that are involved and the strain or species of bacterium under study. Rising entropy levels have been used to predict the success of antibiotic therapy for any type of antibiotic in any strain of bacterium, which is far more efficient than current gene panels that rely on specific gene expression profiles in specific micro-organisms (Zhu et al., 2020). The increased amount of disorder in gene expression profiles that is observed in stressed bacteria has been explained in terms of a loss of regulatory influence (‘dysregulation’), which normally coordinates dependencies between genes and produce some degree of order. The loss of such coordination then causes gene expression levels to vary independently and more randomly. In other words, rising levels of entropy seem to signal a regulatory overload, which is predictive of a loss of fitness.

The predictive power of (permutation) entropy or similar measures generalizes well beyond bacteria. It has been used to predict behavioral changes of a large number of different classes of organisms under stressful conditions, including plant species (Sun et al., 2010), fish and other aquatic organisms (Bae and Park, 2014; Egurraun et al., 2014), insects (e.g. Liu et al., 2011), chickens (Maria et al., 2004), quails, rats, pigs and primates, to name but a few (e.g. Asher et al., 2009). Interestingly, increased disorder has been discovered in timeseries of human behavior under stressful conditions. Human inner experience and overt behavior (the mental phenotype) can be measured using experience sampling methodology (ESM): a technique that involves rating multiple phenotypical items several times a day for several weeks or months to produce timeseries. When stress levels increase, typical changes can be observed in such timeseries that involve increased levels of variance, increased amplitudes, increased anticorrelations between opposing mental states (e.g. happiness and sadness), increased temporal autocorrelations and a slow recovery from external perturbations (van de Leemput et al., 2014). Together, these changes signal the phenomenon

of ‘critical slowing down’ (CSD), which is a highly generic state of network systems that are poised on the brink of a ‘tipping point’ (a sudden transition from one state of the system to another). Just before the onset of such phase transitions, the systems starts to show erratic behavior (CSD). CSD is a generic characteristic that can be used as an early warning sign to predict the occurrence of tipping points in non-living as well as in living systems (Veraart et al., 2012; Scheffer et al., 2012). In humans, CSD has been used successfully to predict the onset of a mental disorder (major depression) at least 3 months in advance (van de Leemput et al., 2014). As is evident from their respective definitions, CSD is actually synonymous with a (transition towards a) state of high (permutation) entropy. Entropy is a much more general term, however, which can be quantified from timeseries data using a single parameter instead of three terms or more ( $H = -\ln |M_p|$ , where  $|M_p|$  denotes the determinant of a graphical lasso regularized empirical correlation matrix), or even from a single timepoint (i.e.  $H_{stp} = \ln(\sigma^2)$ , where  $\sigma$  refers to the measured variance in the expression of recorded variables for a single timepoint) (Zhu et al., 2020). In short, a universal stress response can be formulated not by looking at the specifics of regulatory activity in living systems, but rather at the total amount of disorder observed in hierarchical message passing within organisms (as measured e.g. by a bacterial transcriptome or brain activity). Rising levels of entropy have been proposed to result from a loss of ‘regulatory connections’, which normally coordinate (e.g. synchronize) the different elements of the system and produce order (Zhu et al., 2020). High levels of permutation entropy can serve as a generic early warning sign for sudden state transitions reflecting a failure of control, which signal either stagnant growth, disease or the death of an organism (i.e. a loss of homeostasis). We will now examine whether the overt behavior of organisms under high levels of stress shows universal changes as well, in order to derive a general theory of stress and the stress response in living control systems.

When studying the overt behavior of organisms under high levels of prolonged stress, features emerge that appear universal to all organisms. Whereas short or sublethal stress levels seem to speed up metabolism, promote motility (fight or flight) and enhance exploration tendencies (migration), social activity (establishing hierarchy), the exchange of genetic material (procreation) and parental investment in a wide range of organisms, prolonged and (near) lethal stress levels induce behavioral changes that involve a down-regulation of metabolism (e.g. bacterial stasis, sporulation, hibernation), reduced motility (mobility or migration), reduced sociability, a halt on reproductive activity, an increase in (DNA) repair activity or sleep, and a tendency to neglect (abandon, or even eat) offspring (Wingfield et al., 1998, 2003; Ruf and Geiser, 2015; Hausfater and Hrdy, 2017; Del Giudice, 2020). Such ‘emergency life history stage responses’ generally economize on long-term, (pro)social and/or reproductive activities in favor of short-term, self-repairing and self-preserving activities. In more concise terms, severe stress is said to cause organisms to shift away from ‘slow’ policies (i.e. long-term prosocial activities and parental investment) and towards ‘fast’ policies (i.e. short-term and self-preserving activities) (Del Giudice, 2020). Such shifts in behavioral policies are especially evident in social species (when compared to solitary species), since these normally devote a significant amount of their time in building social hierarchies and parental investment (Del Giudice et al., 2015). Nevertheless, even bacteria are known to cut down on ‘social’ and reproductive activities in response to a (near) lethal stressor, e.g. when shutting down horizontal gene transfer, halting cell division or engaging in sporulation (Lyon, 2015; Meeske et al., 2016). It therefore seems that organisms upregulate complex behavioral policies under intermediate levels of stress but abandon such policies when stress levels approach near lethal levels. Such behavioral changes have been explained in terms of ‘allostatic overload’, which refers to the situation where the regulatory capacity of a control system is overtaxed by environmental perturbations, i.e. where regulatory work increases to the point where energy demand exceeds energy supply (Wingfield et al., 1998, 2003; McEwen and Wingfield,

2003). In such cases, organisms need to cut down on computationally expensive regulatory activities in order to save energy and resources. Interestingly, scholars have linked the expenditure of energy and resources to the hierarchical depth of information processing (Hermans et al., 2014; Goelzer and Fromion, 2017; McEwen and Wingfield, 2003): higher ('allostatic') levels of hierarchical control systems that inspire more complex forms of behavior demand more energy, whereas lower ('homeostatic') levels that control relatively simple behavior require less regulatory work and demand less energy (see introduction). Thus, organisms appear to abandon hierarchically higher levels of processing in favor of lower processing levels when confronted with allostatic (higher regulatory) overload (Hermans et al., 2014). This has been demonstrated experimentally in different organisms including humans, e.g. with human behavior falling back from goal-directed to habitual behavior under severe levels of stress (Schwabe and Wolf, 2011; Goelzer and Fromion, 2017; Van Oort et al., 2017). In the previous section, we saw that self models, social models and transcendent (normative) models involve the highest levels of contextual integration (in a spatial/social and temporal sense) and inspire complex forms of goal-directed behavior, including moral decision making. Consistent with the theory of allostatic overload, severe stress is known to negatively affect prosocial behavior and moral decision making (although moderate levels of stress may actually increase prosocial behavior, see below) (Lee and Yun, 2019; Mendez, 2009; Starcke et al., 2011; Youssef et al., 2012). Extreme stress therefore seems to affect policy selection as a function of contextual integration, i.e. organisms take lesser amounts of contextual information into account when formulating a response. Such decontextualization allows them to revert to more basic policies that demand less energy. In short, increased levels of entropy in hierarchical message passing within living systems are already suggestive of a loss of integrative control under severe levels of stress, but this notion is further backed by studies of the overt behavior of stressed organisms that independently point towards a reduction in model complexity and corresponding shifts in policy selection. We therefore propose that extreme stress causes a top-down collapse of goal hierarchies, i.e. a loss of hierarchical depth. This forces organisms to downgrade from high-level (functionally integrated) goal states to lower-level (functionally segregated) goals and corresponding behavioral policies. The universal presence of this principle suggests it has cornerstone value in securing survival 'when the going gets tough'.

For a mechanistic account on how this might work, it is worthwhile to examine the biophysics of stress in hierarchical (Bayesian) control systems. In such systems, prediction errors are used in two distinct ways (Fig. 6): to update Bayesian beliefs by resetting priors (changing your mind) and to induce an output sequence or stress response to reduce the error via the environment (changing your actions). Lower level policies (e.g. walking) are allowed to run freely until prediction errors are produced within the input hierarchy (e.g. stumbling across some unforeseen object; Scafetta et al., 2009). If error signals cannot be sufficiently suppressed by a simple, straightforward response generated at some lower level of the hierarchy (e.g. side-stepping), the residual error is 'escalated upward' into the hierarchy to update more comprehensive world models and produce a corresponding, more complex output (e.g. walking around the object; de Kleijn et al., 2014d). Thus, prediction errors pass a hierarchical succession of goal states (increasingly complex generative models) and corresponding behavioral strategies until they are suppressed. Such elaborate strategies may eventually suppress prediction errors in ways that simpler forms of behavior cannot (e.g. by successfully walking across the object, reaching the top of a fruit tree, or climbing a social hierarchy). This may explain why intermediate levels of stress initially cause organisms to display more complex behavioral strategies. The vertical accumulation of prediction error can be thought of in terms of a loss of control over free energy. In the Bayesian inference literature, rising levels of free energy are usually associated with increases in entropy and a concomitant loss of thermodynamic and computational efficiency. The use of higher order (more complex)

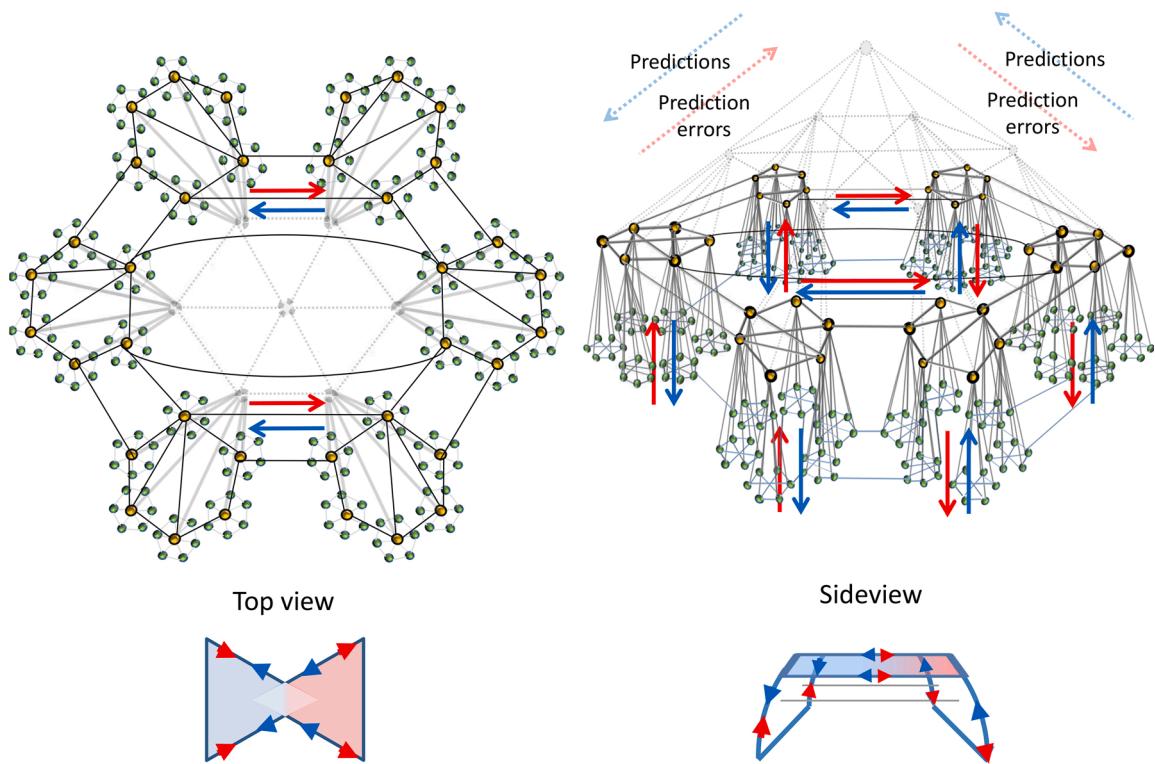
strategies is therefore likely to coincide with rising levels of (permutation) entropy in measures of hierarchical message passing and overt behavior.

Since any hierarchy of control systems is finite, however, prediction error signals may accumulate upwards across multiple levels of control until the top of the hierarchy is reached. At that point, the organism has exhausted its hierarchy of goal states and corresponding policies (i.e. even complex strategies are ineffective at suppressing prediction errors). In such cases, vertically accumulated prediction errors activate a small number of hub structures located near the top of the (goal) hierarchy (the knot of the bow-tie). These hubs maintain many long-distance connections with other network clusters and subclusters in the network, thus representing the highest level of integration within the network structure at large (Figs. 5–7). From simulation studies in statistical physics, it is known that the highest degree nodes in a network have the highest levels of energy dissipation, corresponding to highest energy demand (Gosak et al., 2015). The fruitless pursuit of high-level goal states and corresponding behavioral policies may therefore cause these structures to be flooded with ascending prediction error signals, to the point where energy demand exceeds energy supply. When this happens, these high-level hub structures will overload and fail (Gosak et al., 2015; Stam, 2014). This 'allostatic overload' has been experimentally confirmed to coincide with increased levels of permutation entropy specifically for such hub nodes (Sun et al., 2010). Since high-level hub structures normally integrate information streams across a large number of subordinate clusters (functional integration), their shutdown causes a shift in the balance between functional integration and segregation of network clusters in favor of functional segregation (Tononi et al., 1994). This corresponds to a collapse of the nested modular goal hierarchy: more encompassing goal states effectively 'decompose' into their constituent components, inducing a corresponding change in behavior. As a result, the subordinate network clusters will no longer be functionally connected and start to operate independently (functional segregation). This loss of integrative control and the ensuing uncoordinated activity of subordinate modules will add significantly to increased scores on (permutation) entropy.

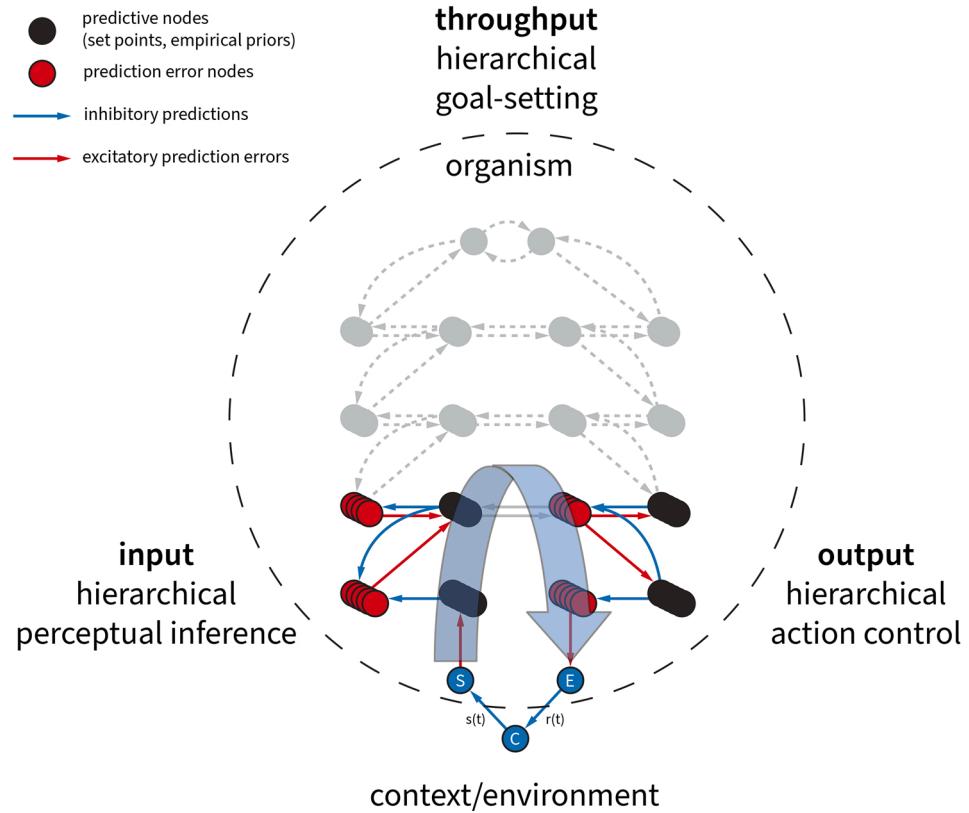
To be clear, we wish to emphasize that hub overload and failure is most likely to involve a *functional* shutdown of high-level predictive hub structures and decreased functional connectivity, rather than a structural loss of hubs and a loss of *structural* connectivity (i.e. failing hub structures still remain physically in place). At least, this seems to be the case in acute forms of stress. In chronic forms of stress, studies show that even a loss of structural connectivity may occur (e.g. synaptic pruning), which may involve the active degradation of maladaptive world models (goal states and corresponding policies) by means of substances such as glucocorticoids (Peters et al., 2017). We predict that the collapse of goal hierarchies is a function of node degree: the most integrative goal states are the first to go, but subordinate levels with lesser-degree hubs and corresponding subgoals may follow depending on the amount of accumulated stress (prediction error). Severe stress may therefore cause a graded disintegration of a nested hierarchy of goal states across several levels. Like military command collapsing in a top-down fashion (generals first, then colonels, lieutenants, higher officers, etcetera), allostatic overload may dissolve goal hierarchies, leaving only the local troops and the odd sergeant major to take care of the problem (Fig. 8). This may explain why severe levels of stress eventually cause organisms to display increasingly primitive forms of behavior. This hypothesis can be tested by examining measures of hierarchical depth of (functional) network structures in relation to policy selection and behavioral complexity at different levels of stress (see Discussion).

When higher-level hub structures overload and fail, they lose their influence as empirical priors that are important in maintaining the balance between top-down prior beliefs and bottom-up sensory evidence (Fig. 6). The overall *amplitude* of prior signals or prediction errors is often quantified in terms of 'precision', which refers to the inverse variability (dispersion) of a probability distribution. In other words,

A.



B.



(caption on next page)

**Fig. 8.** Severe Stress in Organisms: The Collapse of a Hierarchy of Goal States.

**Note:** Hierarchical Bayesian control systems allow organisms to incorporate an increasing number of contextual cues from their environment and create a hierarchy of ‘world models’, i.e. goal states that are used to inform behavior. At the highest levels of integration, such goal hierarchies involve internal (self) models, external (social) models and cross-cutting (normative) models (Fig. 7). In severely stressed organisms, this goal hierarchy collapses in a top-down manner, possibly as a result of hub overload and failure (grey nodes). This results in a ‘decontextualization’ of behavior, with organisms favoring short-term and self-centered policies (informed by self models) over long-term social and/or normative behavior (social and normative models), to save energy and resources. The regulatory collapse may involve several hierarchical levels of integration, depending on the error levels that are encountered. Phenotypically, this manifests as organisms ‘downgrading’ from goal-directed to instrumental, habitual or even reflexive forms of behavior. The top-down loss of hierarchical control by high-level (inhibitory) empirical priors produces both a disinhibition and loss of coordination of lower levels, adding to permutation entropy levels. Please note that individual differences may cause some organisms to retain top-down control under severe levels of stress. See text for further details.

precision scores the ‘confidence’ in a Bayesian belief or prediction error. Increasing the gain of prediction units means that these priors are selected, causing them to more strongly suppress prediction errors and making them more resistant to belief updating. Conversely, increasing the connective efficacy (amplitude or ‘gain’) of prediction error units means that the associated prediction errors are selected, enabling them to preferentially induce belief updating higher in the hierarchy. This leads to the notion of precision at different levels of the hierarchy, whose balance is crucial for determining the relative influence of top-down prior beliefs relative to bottom-up sensory evidence. We can therefore think of stress as reducing **prior** precision and rendering the organism more exposed to belief updating based upon immediate sensory evidence: stress alters the connective efficacy of priors and prediction error units, thereby sequestering them from other levels of the hierarchy. This collapse can either be reversible, e.g. reflecting modulatory control of connection strengths in acutely stressful situations. Alternatively, it could be mediated by long-term changes in connective efficacy or a loss of connections per se, which have been reported in neural systems after chronic stress<sup>3</sup> (e.g. McEwen et al., 2015).

In short, stress seems to change behavioral policies according to a hierarchical principle, i.e. increasingly less contextual cues are used to inform behavior, suggesting a top-down collapse of goal hierarchies. This ‘decontextualization of behavior’ has several short-term advantages. First, organisms will spend less energy and resources on reaching long-term and complicated goals, which allows them to endure current unfavorable conditions for longer periods of time. Second, the bypassing of higher-level systems reduces the path length of the network, allowing signals to travel from input to output structures across shorter distances, producing faster responses (computationally, this is equivalent to minimizing model complexity or computational complexity costs). Third, the top-down collapse of integrative control reduces the gain of predictive connections that normally constrain (inhibit) lower-level policies. This allows such policies to be expressed more freely, making them more pronounced and easy to trigger. This is referred to as ‘disinhibition’ in the psychological sciences and involves a heightening of the senses (within the input hierarchy) and a strengthening of responses (within the output hierarchy), to produce a ‘livening of the reflexes’ (Gorenstein and Newman, 1980). Thus, organisms capitalize on model complexity and precision to formulate a stronger and faster response. This may provide organisms with just the edge needed to force themselves a way out of a dire situation (Byrd et al., 2019).

Such changes come at a price, however, which is a loss of regulatory finesse. A reduction of model complexity makes organisms more vulnerable to environmental conditions that require a broader (and/or long-term) perspective. Additionally, a deep collapse of a regulatory hierarchy may lead to a state of disinhibition where any input almost

immediately triggers a strong output and vice versa. In such a case, even a small environmental disturbance may trigger an intense, volatile, and uncoordinated response (Byrd et al., 2019). This response may then change the environment of the system to the effect that it serves as a trigger for a novel response, and so on. The self-sustaining (circularly causal) pattern of reflexive activity that thus emerges is called a ‘clonus’. This refers to situations where a loss of higher-order inhibitory constraint causes the input and output elements of a control system to become strongly coupled (i.e. the intrinsic coupling of the system is enhanced, causing it to become strongly reactive to input). Such a strong intrinsic coupling then induces a stronger extrinsic coupling (of the organism with its environment) and clonic activity. At some point, the system may become so strongly coupled to its environment that it will lose its ability to compensate for environmental disturbances: it will decompensate (lose control), after which the interior state of the system will linearly follow that of the environment (i.e. a loss of homeostasis). In living systems, such tipping points amounts either to disease, or the death of the organism.

To our knowledge, this is the first detailed model of allostatic overload, or the way in which stress may cause a top-down collapse of high-level integrative control that leads to increased levels of disorder (entropy) in hierarchical message passing and overt behavior in living systems, to eventually produce tipping points (disease or death). Such tipping points occur when such a collapse reaches too deeply down a hierarchy of control systems (i.e. when a hierarchical tree is pruned beyond a level of adequate control). According to this model, organisms may differ in their susceptibility to tipping points as a result of individual differences in the outgrowth (maturation) of their regulatory hierarchies, i.e. different heights of the regulatory tree come with different thresholds for tipping points (decompensation) and, hence, biological fitness. This hypothesis can be tested e.g. by examining the degree to which measures of the hierarchical depth of biological networks predict entropy levels and tipping points under varying levels of stress (see Discussion).

Previously, scholars have defined stress specifically as a failure of control (e.g. Del Giudice et al., 2018), but provided no clear mechanism. Others focused more on physiological states (McEwen and Wingfield, 2003) or cognitive processes in humans (Koolhaas et al., 2011; Ursin and Eriksen, 2010). Most previous definitions of stress situate that state somewhere in between criticality and tipping points as defined above. Here, we employed a more liberal definition of stress as the (cumulative) error state of hierarchical control systems (Peters et al., 2017). The advantage of this definition is that it can be generalized across species and that it lies on a continuum, with clear and objectifiable stress-responses marking discrete levels of stress, i.e. (0) Routine performance (low levels of prediction error, low entropy, reflexive, instinctive (Pavlovian) or habitual behavior, ‘homeostatic control’), (1) Creative problem solving (upward escalation of prediction error signals, rising entropy, goal-directed action, ‘allostatic control’), (2) Emergency responses (high levels of prediction error, high entropy, top-down collapse of goal hierarchies, ‘allostatic overload’, downgrading from goal-directed to lower forms of associative learning, ‘regression to homeostatic control’), (3) Critical slowing down (high prediction error, high entropy, near loss of control) (4) Tipping points (decompensation, loss of control). For a similar categorization of the stress response, see

<sup>3</sup> Neurophysiologically, precision is usually thought to be mediated by the control of synaptic efficacy; either through neuromodulatory transmitter systems or the nonlinear dynamics that mediate synchronous gain. This will be particularly relevant later when we talk about psychopathology. This follows because most of the drugs used in psychiatry act upon the neurotransmitters that modulate synaptic gain and therefore control the precision of message passing in the human brain.

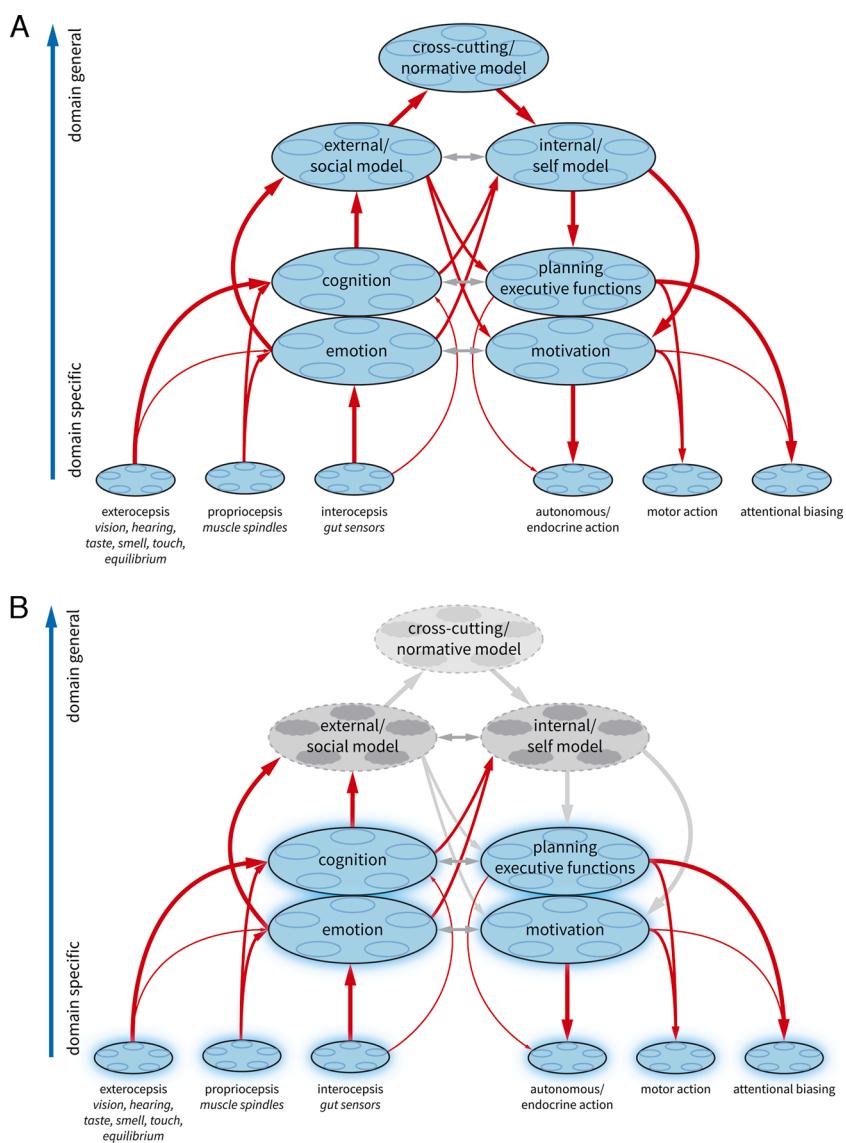
(Romero et al., 2009).

#### 4. The human brain as a hierarchical Bayesian control system

So far, we have discussed rules of network structure and function that may apply to all living systems. We will now show that such rules apply to human behavior. At larger spatial scales, the human brain has a multimodal, hierarchically controlled *small world* network structure (Bullmore and Sporns, 2009; van den Heuvel et al., 2008v). Its 86 billion neurons (Azevedo et al., 2009) form neural modules that are an average of around 5 degrees of separation apart from any other module in the brain (Bassett and Bullmore, 2006; Hilgetag and Goulas, 2016; Sporns and Zwi, 2004; van den Heuvel et al., 2008v). These modules form a nested hierarchy of part-whole relationships (Meunier et al., 2010, 2009) that give rise to a bow-tie network architecture (Markov et al., 2013). Perceptive areas form the input hierarchy of this bow tie, the medial (pre)frontal lobe and anterior insula its knot and (pre)motor cortices and hypothalamic areas make up the output hierarchy. Overt behavior reflects the concerted action of large numbers of simple

input-output patterns at the bottom of this hierarchy ('reflexes', which tie basic input to motor and endocrine output primitives), the activity of which is carefully orchestrated by higher levels of integration (Fig. 8; Botvinick, 2008; Botvinick and Weinstein, 2014; Freeman, 2001, 2005; Ribas-Fernandes et al., 2011). The human brain has been compared to a Bayesian inference engine, whose primary job it is to infer the (hidden) causes of its sensory input by building predictive models of the world and acting upon those models (Friston, 2010; Friston et al., 2006). In doing so, a generative model is constructed with multiple hierarchical levels of model complexity that constitutes our inner experience and overt behavior (the human mental phenotype, or 'mind'). Fig. 9 summarizes current ideas on the human mind as a hierarchical generative model that has its origin in different forms of perceptive information (Badcock et al., 2019; see below for further references). The statistical structure of this phenotypical hierarchy is assumed to mirror that of the human brain (i.e. it has a nested modular bow-tie network structure).

At the bottom of the phenotypical hierarchy, three global types of perceptual input can be discerned. Exteroceptive perception involves information coming from the external environment, i.e. the main senses



**Fig. 9.** The Human Mind as a Hierarchical Generative (Bayesian) Model.

**Note:** A. The human mind can be modeled as a nested modular hierarchical generative model that controls perception and action (Badcock et al., 2019). This figure summarizes current ideas on the statistical dependencies between the different components of this hierarchy, which are assumed to echo those of human brain function (i.e. a nested modular folded bow-tie structure). Circles indicate generative models and circles within circles subordinate models. Higher level (domain general) models are inferred from progressively lower (domain specific) models that eventually have their origin in different forms of perceptive information (e.g. exteroceptive, proprioceptive and interoceptive domains). Arrows sizes reflect the putative contribution of a particular domain in biasing inference within another domain (see text for further details). Note that cognition, emotion, executive functions and motivation occupy a similar hierarchical level of inference (the Figure is 3D). Domains may affect each other across loops of different pathlengths (e.g. from input to output via a hierarchically ordered set of goal states), reflecting different degrees of information processing (policy selection). The shortest loops within this hierarchy represent basic stimulus-response patterns (e.g. simple and more complex 'reflexes', Pavlovian instinct patterns and habitual behavior), whereas the longest loops reflect goal-directed behavior that is informed by highly integrated world models involving self-referential, social and normative models (Fig. 6). Each phenotypical domain may have multiple functional-anatomical brain regions as a correlate (see text and references for further details). B. When stressed severely, contextually redundant higher-level goal states are shut down to save energy and to enhance the stress response (Fig. 8). This corresponds to a collapse of self, social and/or normative models, causing a shift away from goal-directed behavior (longer loops) towards habitual, instinctive or reflexive behavior (progressively shorter loops). The loss of higher-level integrative constraint triggers a disinhibited and disordered state at lower levels within the hierarchy (glow), involving emotional, motivational, cognitive, perceptive, premotor executive and action domains. A shallow collapse may provide leverage out of a difficult situation, but a deep collapse will cause the current model to revert to a hierarchical model of psychopathology (HiToP, see text). In such cases, the system will show increased intrinsic connectivity, which enhances the extrinsic connectivity of the individual, i.e. an increased dependence on the environment and decreased homeostasis. This may present e.g. as strong interpersonal dependencies and/or social conflict. When goal hierarchies fail to mature, such underregulated states become chronic (e.g. personality disorders). See text for details.

of vision, hearing, touch, smell and taste. Interoceptive information involves information feeds coming from the internal environment, e.g. gut and vascular pain and blood pressure afferents, blood glucose concentrations, smooth muscle tension, et cetera. Finally, proprioceptive information takes up position in between the internal and external environment and mostly involves input from skeletal (striated) muscles, tendons and bones. It is thought that hierarchical inference on these basic input domains progressively produces the human mind (Badcock et al., 2019, Fig. 9). Recent studies conceptualize human emotion as hierarchical Bayesian inference on predominantly interoceptive information, placing this hierarchy of affective generative models somewhere along the middle of the larger hierarchy (Seth and Friston, 2016; Smith et al., 2019a). Similarly, cognition may involve hierarchical inference on predominantly exteroceptive information (Smith et al., 2019b). Executive functions in turn involve part of an output hierarchy that is engaged in high-level (conceptual and premotor) planning, with a predominant connection to motor output (controlling muscle action) (Pezzulo, 2012). Motivational functions have been conceptualized as aiding in predicting the precision of motor and endocrine output, with a possible emphasis on endocrine action (Pezzulo et al., 2018). As discussed above, the top of this hierarchy involves highly integrated generative models of their inner (self) and outer (other) states, along with their histories and possible futures. Self models are processed in midline areas of the human brain, which are involved in some of the highest levels of integrative processing (Haggard, 2017; Northoff et al., 2006; van der Meer et al., 2010v; Thornton et al., 2019). Additionally, humans make highly integrated models of the states of others and their possible histories and futures. Such social models (or ‘theories of mind’) involve medioprefrontal, (superior) temporal and temporoparietal areas (Amadio and Frith, 2006; Gallagher and Frith, 2003; Mars et al., 2013; Thornton et al., 2019), which process information at very high levels of contextual integration. Finally, a large body of literature has identified brain regions that are involved in making decisions about events that go beyond the self or the immediate social environment, but instead involve common social laws and values. These normative structures include ventromedial areas for norm processing and right insula, dorsolateral prefrontal, and dorsal cingulate cortices for processing in relation to social norm violation (Zinchenko and Arsalidou, 2018). Such brain areas again involve some of the highest levels of integration across subordinate systems. Together, such studies provide both phenotypical and neuroanatomical support for the existence of a hierarchy of generative models with interior (self), exterior (social) and normative structures at the top of this hierarchy.

Overall, our brains seem to have capitalized particularly on information processing at high levels of functional integration, making detailed predictions of events that take place more distally in time as well as in (interpersonal) space (Herrmann et al., 2007). The ability of the human brain to take large amounts of contextual information into account when formulating a response seems to explain much of its disproportionate size (Dunbar and Shultz, 2007). Despite such extensions, however, the basic principles of control theory that govern behavior in lower organisms remain the same as in humans. As in woodlice, activity levels drop (i.e. we become quiet and pleased) when the perception of our past, current and future environment agrees with our intricate interpersonal goals and expectations.

When observing human brain function and behavior under severe levels of stress, several things stand out. Although mild forms of stress differentially affect or even enhance our personal sense of identity, promote social cohesion or a sense of global connectedness, severe stress brings us into ‘survival mode’ (Buchanan and Preston, 2014; Mao et al., 2016; McEwen and Wingfield, 2003; Von Dawans et al., 2012). Neuroimaging studies show that the human brain falls back from goal-directed to habitual control during severe stress (Schwabe and Wolf, 2009, 2011). This corresponds to decreased activity in higher level systems such as the anterior cingulate, anterolateral insular and temporopolar areas (Arnsten, 2009; Dias-Ferreira et al., 2009; McEwen et al., 2015;

McTeague et al., 2016; Schwabe and Wolf, 2009, 2011; Van Oort et al., 2017). Brain areas that decrease activity during severe stress are midline structures involved in generating self-models (self-image; Goette et al., 2015; Hooley et al., 2005; Kesting et al., 2013; Staniloiu and Markowitzch, 2012), as well as brain areas associated with the production of social world models or theory of mind (Sandi and Haller, 2015; Todd et al., 2015), producing more selfish forms of behavior. Finally, severe stress is known to negatively affect moral decision making (Lee and Yun, 2019; Mendez, 2009; Starcke et al., 2011; Youssef et al., 2012). This change in behavior is related to altered activity in brain areas involved in transpersonal identification, including law-abiding and moral behavior (Lee and Yun, 2019). Thus, severe stress decreases activity specifically within brain areas that support some of the highest forms of contextual integration. Such findings support the hypothesis that significant stress causes a top-down collapse of deep goal hierarchies to save energy and resources, causing people to take increasingly less amounts of contextual information into account when formulating a response (Figs. 8 and 9). Of course, individual differences may cause some people to deviate from this general pattern.

In short, we propose that severe stress prunes the top of a regulatory pyramid in people’s brains to produce a subtle form of decortication (hypofrontality, or a ‘chicken without a head’ syndrome). Such a top-down collapse of goal hierarchies reduces the amount of integrative control (lowers the gain of inhibitory empirical priors), which increases disorder at subordinate levels of the hierarchy, down to the level of the shortest reflex loops. This may manifest as a more violent expression of behavioral primitives or stress response patterns such as fight, flight, fright, feeding, freezing, reproducing, fainting, fawning, etcetera. This disorganized state may underlie increased levels of entropy observed in the overt behavior of severely stressed subjects, which are known to predict the onset of tipping points. Such ‘decompensation’ or ‘dysregulation’ can serve as a generic model for episodic mental illness (van de Leemput et al., 2014). In such cases, the hierarchical generative model as shown in Fig. 9B reverts to a hierarchical taxonomy of episodic mental illness (‘psychopathology’) (HiToP - Kotov et al., 2017). The relative contribution of each phenotypical domain to the overall disease presentation can be parsimoniously expressed as a transdiagnostic factor profile. A differential collapse of the world models of self-functioning, interpersonal functioning and normative functioning should then be a common factor in all forms of mental illness (whether episodic or chronic). The specific type of episodic mental disorder is then determined by the subordinate modules and behavioral primitives that show (disinhibited) disorder as a result of losing these highest levels of integrative control.

This idea is supported by phenotypical studies that show decreased scores on measures of self-functioning and interpersonal functioning as common factors in different forms of mental illness (e.g. Sleep et al., 2019). Also, recent findings show that some changes in brain function are common to a diverse range of acute mental disorders (e.g. major unipolar depression, bipolar disorders, psychosis and anxiety disorders). Such disorders are accompanied by ‘transdiagnostic changes’ in functional neuroanatomy, which include decreased activity levels in prefrontal and anterior brain regions that support high-level cognitive control (McTeague et al., 2016). These are the same areas that harbor our world models of self, others and global world views (Brunner et al., 2010), which are downregulated under stress (see above). Together, such findings support the idea that all forms of episodic mental illness involve a temporary collapse of higher levels of control that reaches too deeply down the hierarchy. This hypothesis can be tested by studying measures of hierarchical depth in different brain areas as a function of entropy (disorder) levels in hierarchical message passing and corresponding phenotypical changes in healthy controls and patients with different forms of mental illness (see Discussion).

From the above, it follows that individual differences in the degree to which goal hierarchies have grown and matured in the course of life should explain different susceptibilities to mental illness (disorder and

tipping points): people with strongly matured hierarchical trees may better withstand the pruning of their hierarchies during a stressful episode than people with lesser developed hierarchical trees. Interestingly, the development of goal hierarchies across the lifespan has been linked to personality development (Russell Cropanzano and Citera, 1993). This process involves the outgrowth and sculpting of goal hierarchies as a result of different forms of associative learning of organisms in relation to themselves and their environments ('maturation'), see above. Whereas episodic mental disorders involve a temporary collapse of goal hierarchies, personality deficits may involve a failure of such structures to develop normally. Neuroimaging studies show that people with (borderline) personality disorders, who are more susceptible to mental decompensation ('crises'), have low volumes of gray matter in the same areas of high-level (cognitive) control that are downregulated under stress (Brunner et al., 2010; McTeague et al., 2016). These underdeveloped brain areas involve the same areas that harbor our world models of self, others and global world views (Brunner et al., 2010). The global faculty of cognitive control that is down-regulated in acute mental illness can therefore be subdivided into high-level world models that support self-functioning (agency), interpersonal functioning (communion) and normative models (meaning), which may each be downregulated to different degrees under stress (Fig. 9B). These mental faculties therefore qualify as 'transdiagnostic factors', which are to some degree involved in all personality disorders (when underdeveloped) and episodic mental disorders (when downregulated).

The collapse of these transdiagnostic world models under stress may cause people to experience specific sets of 'symptoms', i.e. a decreased sense of purpose and normativity (due to collapsing normative functions), a loss of empathic interest in others or the external world, a decreased feeling of communion, derealization (due to a collapse of social / external models), or rather undecisiveness, low self esteem, distorted body-image and/or symptoms such as depersonalization and dissociation (a clinical state characterized by a loss of self-awareness caused by a collapse of self-models). These are typical symptoms of (borderline) patients during acute episodes and may to some degree be common to all patients with mental illness. Indeed, the 'alternative model' for personality disorders in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) currently lists self-referential and interpersonal functions as two global transdiagnostic factors that are underdeveloped in personality disorders (Zimmermann et al., 2015). These may at some point be supplemented with the third overarching factor (normative functions) as identified in the current paper, a conclusion that is consistent with some existing models of personality development (e.g. Cloninger, 2008). The maturation of these 'great three' world models involves a life-long process of goal-directed learning. The development of these mental domains across the lifespan has been referred to as personality development (or more specifically 'character' formation) (Cloninger, 2008). The various generative models that are subordinate to these three top-level domains (i.e. emotional, motivational, cognitive and executive domains and subdomains) qualify as lesser-order transdiagnostic factors (Fig. 9A). These involve shorter stimulus-response loops that have been associated with Pavlovian learning and habit learning. Such functions develop at earlier stages of life and their stable expression across years has been referred to as 'temperament' (Cloninger, 2008). Individual differences in the expression of such factors are known to produce different personality profiles and susceptibilities for episodic disorders. Together, such findings support the idea that a stress-induced collapse of already underdeveloped regulatory hierarchies triggers disorder and tipping points in human subjects with personality disorders, with shallower hierarchies increasing the risk of such episodes. Future studies may link the hierarchical depth of regulatory hierarchies to scores on specific personality domains and susceptibilities to episodic disorders.

Apart from explaining individual differences in behavior and (susceptibility to) mental illness, the current model may explain individual differences in social interactions. This is because individual organisms

can be modeled as hierarchical Bayesian control systems that respond to each other, i.e. the output of one individual (behavior) may serve as the input to another (Friston and Frith, 2015). Such models allow for studies on interpersonal dynamics at small timescales (e.g. stress-induced changes) or at larger timescales (e.g. developmental differences). For instance, a top-down collapse of higher order control may increase extrinsic (social) coupling of one individual with respect to another. This may then cause a collapse of higher order control in the other person (e.g. through a lack of sleep), producing highly recursive (clonic) stimulus-response relationships between two individuals. As a result, two undercontrolled (stressed) individuals may become strongly coupled. This would be a model of strong mutual dependency and/or intense social conflict, including a mutual loss of law-abiding and moral behavior. Much like clonic spinal reflexes that can be silenced only by an external influence, vicious cycles in social behavior are a symptom of insufficient higher-level control that typically require an external party (e.g. mediation, judicial arbitrariness, or medical intervention) in order to be reduced (Fehr and Fischbacher, 2004).

Thus, individual differences in hierarchical Bayesian control (e.g. personality development) produce stable differences in social interaction, which translate into stereotypical connectivity patterns at a local level, e.g. scores on personality dimensions predict the topological position of individuals in social networks (e.g. Krause et al., 2010). Such individual differences in local connectivity act as simple rules that knit together complex social networks at a global level. This includes the formation of social network clusters in which some opinions and beliefs are held and contrasted with those of other individuals or groups, while trying to get a mutual grip on reality. Social networks may therefore follow similar rules for network architecture and function (collective inference) as shown in Fig. 6.

## 5. Discussion

In the current paper, we present a universal theory on information processing in living systems as well as a general theory on stress and the stress response that are based on first principles in biophysics. We propose that all living systems can be modeled as scale free, *small world* (nested modular) network structures with an information bottleneck structure, resulting in hierarchically organized input (perception), throughput (goal setting) and output (action) parts that are engaged in Bayesian inference. To our knowledge, this is the first time that concepts from network science and graph theory are put together with current ideas on predictive coding to explain hierarchical Bayesian inference in living systems (e.g. Friston et al. (2017)). When embedded in an (a)biotic environment and allowed some freedom of movement, such systems function as 'hierarchical Bayesian control systems', which change their environments through action in order to reduce the difference (error) between their perception of their current inner or outer state (posterior) and their self-inferred goal states (priors: predictive models of the world of varying levels of complexity). The minimization of this error through either action or model adjustment (learning) corresponds to a gradient search on mean variational free energy, which is known as 'active inference'. Error can be minimized with respect to a hierarchy of goals and corresponding subgoals, with the top of the goal hierarchy representing the most integrated ('highest') goals of the organism. Such goal hierarchies allow organisms to perform an iterative search for ecological niches of a certain predilection, i.e. niche exploration. The prediction error (free energy levels) of hierarchical Bayesian control systems can be defined as 'stress' and the action that follows the error as the 'stress response'. Under stressful conditions, error accumulates vertically in the goal hierarchy and increases the oscillation frequency of network nodes until energy demand exceeds energy supply ('allostatic overload'). The most connected (highest degree, or central) nodes at the top of the information bottleneck (goal hierarchy) are most vulnerable to such energy depletion, causing them to selectively overload and fail. To our knowledge, this is the first explicit mechanistic model of allostatic

overload. The selective loss of central (hub) nodes results in a top-down collapse of goal hierarchies, causing organisms to abandon hierarchically higher (more integrated and abstract) goals in favor of hierarchically lower (less integrated and more concrete) goals to save energy and resources. This corresponds to a shift in behavior from long-term, cooperative (prosocial) and/or selfless (altruistic) policies towards short-term, solitary (asocial) and/or self-centered (antisocial) policies. In humans, the (hierarchically) highest goals correspond to social norms and moral values that individuals deem applicable across living systems and timescales. The collapse of such goal states and corresponding behavioral changes under stress corresponds to a blunting of social interactions and, eventually, moral decay (of course, individual differences may cause subjects to deviate from this general rule). Studies indicate that high levels of stress are accompanied by an increase in permutation entropy in measures of hierarchical message passing and overt behavior. Permutation entropy is a measure of ‘disorder’ in timeseries that quantifies a number of erratic changes. In many living systems, increases in permutation entropy successfully predict a sudden phase transition (a tipping point), indicating disease, or the death of an organism. We propose that an increase in permutation entropy signifies a loss of higher level integrative control across lower-order systems, causing these systems to behave in an uncoordinated, desynchronized (erratic, disordered) way. In humans, a temporary collapse of high-level (integrative) goal states may underlie episodic mental disorders (e.g. major depression, psychosis, panic attacks), whereas a failure of goal hierarchies to mature in the course of life may serve as a model of personality (trait) disorders. The term ‘disorder’ therefore seems well-chosen, since it points out an increase in permutation entropy. Whereas a shallow collapse may only cause small changes in behavior, a deeper collapse of goal hierarchies diminishes model complexity to the point where stimuli almost immediately trigger (stress)responses and circularly causal (clonic) patterns emerge between the organism and its environment. Such vicious cycles or attractor states signal a loss of homeostasis and usually align with disease, or the death of an organism. Such changes are universal features of living systems and can be observed at any scale level of organization, including social levels. In order to test these predictions, researchers may need to consider the whole of hierarchical message passing in organisms instead of just parts of it. This has been a major obstacle in the past, but modern data analysis techniques increasingly allow studies of the full complexity of interactions between genes, proteins, metabolites, neurons, brain areas, phenotypes, animal populations and people (the -omics literature). Below, we will discuss several ways of testing these predictions.

### 5.1. General architecture

Our first prediction follows from the universal presence of *small world* topologies in living systems (see Introduction). As a result of this universality, we expect living network systems of any type to show commonalities in network structure. Network structure can be analyzed using software packages such as the igraph library in R (Csardi and Nepusz, 2005) or Cytoscape (Shannon et al., 2003). Small-worldness can be quantified by calculating a small-worldness index, which compares the clustering coefficient (modularity) and average path length of given network to a randomly connected network of equal size (Humphries and Gurney, 2008). A value significantly greater than 1 (and preferably more) indicates that the network is non-randomly connected and contains hub nodes and clusters that allow energy to dissipate along short and efficient paths. Hub nodes can be identified by examining the degree (number of connections) per node, and centrality measures can be calculated that examine the relative importance of nodes in guiding traffic across a network. Hub structures contract their neighboring nodes into network clusters, which can be detected quantitatively by means of network community detection algorithms (e.g. Newman, 2004). Software has been developed that allows detection of so called ‘rich club’ structures (Fig. 4), which are collections of hub nodes that connect

significantly more to other hubs than chance levels (Opsahl et al., 2008). Rich clubs are nested hierarchies of hub nodes that produce a scale invariant network structure. In such structures, each network cluster can be modeled as a node at a next level of spatial aggregation. Functional integration within nested rich clubs structures is an important ingredient of hierarchical Bayesian inference. Also, software packages exist that can test network structures for a diverse range of motifs, e.g. (Masoudi-Nejad et al., 2012). These include bow-tie motifs as well as their constituent motifs, such as feedforward and feedback loops. At the organism level, we expect biological networks to show a nested bow-tie (bottleneck) structure, with cross-connections between similar levels of input and out hierarchies of a (folded) bow-tie, producing input-throughput-output loops of different path lengths. We expect bow-tie motifs to consist of a family of smaller motifs that include feedforward and feedback loops. Studies have already shown an abundance of the feedforward loop motif, which we expect to reflect top-down predictive processing in input hierarchies and bottom-up predictive processing in output hierarchies (Figure 6, Box 1). Such motifs should be counterbalanced by feedback motifs that reflect bottom-up correction of higher-level predictions by lower level prediction errors in input hierarchies (and vice versa in output hierarchies).

With respect to energy flows across biological network structures (network ‘function’) and its directionality, a distinction can be made between global (macrolevel) and local (microlevel) flows. The input hierarchies of nested bow-tie structures should show multiple excitatory energy streams converging onto higher level hub structures while ascending in the hierarchy, reflecting the functional integration of prediction error signals. Also, input hierarchies should show multiple energy streams diverging while descending in such hierarchy, reflecting top-down and inhibitive predictive control. Together, both information streams reflect perceptive inference. We propose that the directionalities of prediction and prediction error streams are reversed in output hierarchies when compared to input hierarchies, reflecting hierarchical action control (Fig. 6). With respect to local flows, we expect input areas of bow-tie motifs to contain a large proportion of hub nodes with multiple arrows converging onto each hub. Such ‘integrator hubs’ (sinks, or driver hubs; Yan and He, 2011) are said to have a high in-degree, referring to the number of incoming connections from other nodes that indicate the process of functional integration. Conversely, the output areas of bow-tie motifs should contain a significant proportion of network motifs that involve multiple outputs diverging from a single (hub) node onto a distributed set of other nodes. Such ‘distributor hubs’ (sources, or driver hubs; Yan and He, 2011) have a large out-degree, referring to the number of outgoing connections that support the process of action control. The throughput parts (knots of bow-ties) may show a substantial number of sources, sinks, and hubs with balanced numbers of incoming and outgoing connections, reflecting continuous cross-evaluation. The net in- and out-degrees of prediction error or predictive hubs are expected to shift along a gradient from input to throughput and output parts of the network, reflecting a smooth transition between these domains. As observed, we expect the dynamics of bottom-up and top-down units (as well as within-level dynamics) to produce oscillatory behavior of different spatiotemporal scale, i.e. attractor states.

Predictions with respect to the directionality of links in biological networks can be tested using software developed to study causal relationships (conditional dependencies in time) between mutually dependent variables (e.g. Scheines et al., 1998). Inferring directions amongst variables using causal reasoning software is considered a hard problem in statistics and the directions obtained may not always be reliable. In networks in which nodes have clear and measurable relationships (e.g. genomic, proteomic or neural networks), it may be quite feasible to infer directions, whereas in other networks (e.g. statistical networks used to study brain function or phenotypical states), testing these hypotheses may prove to be more difficult. Recent attempts to infer both global (Hillebrand et al., 2016) and local (Märtens et al.,

2017) directionalities of functional connections in the human brain have involved the use of a novel and promising phase transfer entropy measure (Lobier et al., 2014). Interestingly, this measure may partly reflect the flow of prediction error (free energy) across network systems, since entropy and energy measures are related through the second law of thermodynamics. Using phase transfer entropy, a bottom-up convergence was demonstrated in sensory areas, consistent with ascending prediction errors within the first part of a bow-tie structure. Evidence for top-down divergence was less clear, however. At a more local level, a bidirectional convergence/divergence motif was found, which may reflect true bidirectionality or an insufficient decomposition of flow directionalities into ascending (excitatory) prediction errors and descending (inhibitory) predictions. Overall, the quantification of energy flows and their directions within biological networks is an important venue for further study. Similar measures that are used to study directionality of energy flows in brain function can be used to study molecular or neural networks.

Several of the predictions made in this paper require a quantification of the concept of ‘hierarchy’. Despite its common use in everyday language, it has proven a challenge to produce a formal definition of hierarchy, hence several definitions exist (Corominas-Murtra et al., 2013). In *small world* networks, some nodes or clusters may only exist by virtue of other nodes or clusters, i.e. they form conditional dependencies in space (a hierarchy of part–whole relationships; Ravasz and Barabasi, 2003). Additionally, biological networks involve state changes that follow a hierarchy of conditional dependencies in time (i.e. causal order, or directionality). Both hierarchies need to be accounted for in order to obtain an idea of the hierarchical order of nodes or clusters in scale invariant network structures. Perhaps the most formal definition of hierarchy is provided by Corominas-Murtra et al. (2013). The authors propose to quantify hierarchy in terms of three key elements, which include treeness (pyramidal shape, or spatial hierarchy), feedforwardness (top-down or bottom-up directionalities, or temporal hierarchy) and orderability (the effect of causal cycles), allowing the hierarchical structure of different types of networks to be directly compared within a single three-dimensional space. This definition of hierarchy controls for the nestedness and directionality of links, but needs to be adapted for weighted networks. Perhaps a more straight-forward approach to measuring the number of hierarchical levels of a biological network structure would be to count the number of nested relationships between clusters and subclusters (i.e. scale levels) regardless of directionality (Kaiser and Hilgetag, 2010). The number of functionally segregated subclusters that are integrated in a nested fashion into a particular hierarchy of control provides a measure of the height of a hierarchical tree (Newman and Girvan, 2004). Several hierarchical network clustering algorithms exist that can provide information on the number of part–whole relationships, allowing for the construction of corresponding tree-graphs (e.g. Lancichinetti and Fortunato, 2009). Measures of nestedness (hierarchical depth) should be intimately tied to the proportion of functional integration versus segregation of network clusters. This relationship can be tested quantitatively by using another measure derived from neuroscience, called neural complexity (CN; Rubinov and Sporns, 2010; Tononi et al., 1994). This measure defines functional segregation as the relative statistical independence of small clusters of a system and functional integration as significant deviations from independence of larger clusters. CN expresses the average deviation from statistical independence for clusters of increasing size. CN values are high when functional segregation and integration coexist in a balanced manner and low when the components of a system are either completely independent (segregated) or completely dependent (integrated). Although first used to analyze neural networks, this measure captures a universal feature of biological systems (Rubinov and Sporns, 2010). Although CN is a structural measure, it may well serve as a means to quantify Bayesian model complexity, which involves the number of independent variables (degrees of freedom) that are available to a particular model. Model

complexity is expected to decrease when moving up the hierarchy of generative models, since higher level models offer a more parsimonious explanation of lower-order events (Spiegelhalter et al., 2002). Other measures to quantify information integration and corresponding fitness have been suggested as well, e.g. Edlund et al. (2011). Together, these measures of (nested) hierarchical depth and model complexity can be used to test predictions with respect to the comprehensiveness of hierarchical control in biological networks (see previous sections for such predictions). Briefly, we expect the amount of functional integration across multiple contextual cues (and the corresponding height of the nested hierarchical tree) to differ between lower (less) and higher (more) organisms, and individuals or species with lower (less) or higher (more) levels of autonomy/agency and self-directedness, solitary (less) and more social (more) behavior, less (less) and more (more) prosocial behavior, smaller (less) and larger (more) amounts of parental investment, less (less) or more (more) transgenerational awareness and actions, less (less) and more (more) normative (law abiding) or moral behavior, and between calm (more) and stressful (less) situations (see below). Such differences may involve specific parts of the network, e.g. throughput hierarchies may show greater (individual) differences in hierarchical depth than perceptive or output hierarchies.

As discussed, hierarchical depth is related to the ability of an organism to control its internal states or the world around it. Organisms with lesser developed hierarchies may therefore find it more difficult to adapt to complex and changing environments. In the specifically human case, the maturation of deep goal hierarchies in humans can be linked to personality development, and insufficient maturation of hierarchical trees to personality disorders and instability (mental illness). Such deficits eventually decrease scores on measures of self models (agency), social models (communion) and normative models (meaning). Future studies may compare the hierarchical network structure of subjects with and without personality disorders to further test these predictions, e.g. using neuroimaging techniques. As observed in Section 4, individual differences in the height or maturation of goal hierarchies can also be linked to stable individual differences in social interaction, which define the local topology in social networks to eventually affect the global structure of social networks.

As a general remark, hierarchical Bayesian inference describes a mechanism for inferring ‘signs out of signs’, which amounts to a model of semiotics (Fortier and Friedman, 2018). Social connections can be defined in terms of the exchange of free energy between different agents through synchronized action-perception cycles and have produced a novel way of thinking about reciprocity and hermeneutics (Friston and Frith, 2015; Vasil et al., 2020). Organisms may act in such a way as to alter the amount of free energy (model error, stress) in other beings. This corresponds to aiding other organism with information or hampering them by not sharing information or providing desinformation, which has a strong moral connotation. Indeed, our model predicts that organisms and people that produce the most detailed and accurate models of the world are at a thermodynamic disadvantage when operating alone (but at a significant advantage when working together). The current paper sees hierarchical Bayesian inference as a way to explain our highest levels of mental functioning, including the formation of social norms and moral goals. Individuals may differ in the degree to which such models have developed and therefore differ in the degree to which their behavior is guided by higher moral principles. Such topics have been kept to the realms of philosophy for many thousands of years. Especially as regards moral functioning, one should be careful not to commit to a naturalistic fallacy by assuming that the factual structure and dynamics of biological systems automatically informs us of a desirable structure (Moore and Baldwin, 1993). Although one should be prudent, however, it is not impossible to move from facts (‘is’) to moral prescriptions (‘ought’), especially when such facts involve things of a hierarchical generative and symbolic nature (i.e. humans as symbolic animals). The relative autonomy of high-level generative models with respect to the lower-level events from which they have been inferred makes it possible

to produce highly creative models that go well beyond the available facts (predictions, goals), yet still have their basis in such facts (memories). This relative disconnection may be what is required to finally integrate science and morality safely within a single discipline (the ‘moral sciences’; Ruse, 1988). This being said, it may well be a ‘categorical imperative’ for all people to actively develop mature regulatory hierarchies that incorporate as many contextual cues as possible into self-transcending world models that allow our behavior to be informed by universal laws and social standards through which people may connect (predict each other) across nations, cultures and timescales. Our scientific, legal and moral institutions may facilitate the exchange of such commonly held values or goals in order to facilitate belief updating and develop a commonly held world view. The detrimental effects of (chronic) stress on such a development should be actively countered across many generations.

## 5.2. Stressful conditions

Organisms continuously change their wiring patterns while anticipating and responding to different situations. This produces a dynamic balance between the functional segregation and integration of network communities and, therefore, hierarchical structure (Sporns, 2013). In this paper, we propose that severe stress alters network community structure of biological systems in a universal way, i.e. it should produce a shutdown of high-level hub structures within the information bottlenecks of organisms (i.e. the knots of bow ties). This shifts the balance between functional integration and segregation towards functional segregation, thereby reducing the height of the nested hierarchical tree (see above). Such changes may not cause a significant shift in the small-worldness measure, but may decrease hierarchical depth as measured by hierarchical clustering algorithms. Thus, severe stress should produce shorter and shallower bow-tie motifs with wider knots, which interferes with the ability of organisms to compress information. This should translate into increasingly shorter loops that run from input via processing to output parts of a (functional connectivity) network. This can be tested by measuring the path length measure from input to output structures for different nodes of interest (i.e. the average distance from one node to another via a subset of intermediate nodes). We expect measures of hierarchical depth to be high in moderately stressful situations and low under either very low or very high levels of stress (i.e. either complete segregation or integration). Additionally, we expect stress to shift the balance between predictive processing and corrections of such predictions in favor of prediction errors, making organisms more susceptible to belief updating by immediate sensory evidence. Such changes involve an increase in the synchronous gain (precision) of prediction error signals versus predictive signals, which involve changes in connective efficacy e.g. as a result of (neuro)modulatory signaling pathways in neural or molecular networks. The overall result of such changes may be examined by measuring shifts in scores on measures of directed and weighted connectivity, e.g. phase transfer entropy studies showing increase bottom-up convergence as opposed to top down divergence in perceptive or goal hierarchies and vice versa in output hierarchies.

Finally, when stress levels are particularly high, we expect tell-tale signs of undercontrolled control systems in the form of increased permutation entropy (or critical slowing down) and changes in overt behavior that signify a reduction in model complexity. This can be tested by linking entropy levels and tipping point thresholds to measures of hierarchical depth and behavioral changes in different individuals or species. Such studies are readily performed in bacteria and other microbes, where e.g. acidity, salinity or antibiotic levels may be varied to examine bacterial responses in hierarchical message passing and growth or survival rates (Nagar et al., 2016; Marles-Wright et al., 2008; Yu and Gerstein, 2006; Zhu et al., 2020). For obvious reasons, however, such studies cannot be easily translated to higher organisms. Actively bringing sentient creatures to the brink of a tipping point would be

highly unethical. In the specifically human case, severe stress does appear to decrease the amount of functional integration within the human brain, as measured by an information processing efficiency measure (Rubinov and Sporns, 2010; Wheelock et al., 2018). Another study in post-traumatic stress syndrome reports increased amounts of functional segregation (Zhu et al., 2019). Yet other studies show that the human brain falls back from goal-directed to habitual control during stress (Schwabe and Wolf, 2009, 2011). Such findings are in line with a collapse of high-level integrative control, but require a systematic approach in order to prove the principles put forward in this paper. Although experimental studies are precluded, however, studies of mental illness may provide a natural situation in which to examine tipping point thresholds in relation to hierarchical depth in humans. As observed, we expect individual differences in the hierarchical depth of goal hierarchies to explain individual differences in resilience and susceptibility to mental disease. Such hierarchies ultimately involve highly integrated self models, social models and transcendent world models. A temporary collapse of these models should be a common (transdiagnostic) factor in all episodic forms of mental illness ('psychopathology'). Conversely, a persistent failure of similar cortical hierarchies to mature properly should underlie a stagnation of personality development and the concomitant chronic risk of episodic mental illness ('personality pathology'). In terms of diagnostics, monitoring scores of patients on these three global domains therefore seems crucial. In terms of prognostics, measuring change scores on these three global domains may help to predict treatment success and relapse rates, whereas entropy levels in ESM timeseries may help to predict the onset of episodic mental disorders (i.e. tipping points, (van de Leemput et al., 2014)). In terms of therapeutics, promoting an optimal balance between top-down predictions and bottom-up belief updating can be performed by prescribing medication that modulates synaptic gain (e.g. antidepressants, antipsychotics, etcetera) or by means of transdiagnostic psychotherapeutic interventions that promote the updating of false beliefs with respect to self, others, and global world views (e.g. exposure and re-appraisal in cognitive behavioral therapy). Of course, much can be won by prevention strategies that discourage people from developing maladaptive world models in the first place, e.g. by providing children with a safe social environment, proper training and education.

To summarize, we expect stress to alter (functional) connectivity in living systems in canonical ways, regardless of whether that involves single-cellular life forms of complex multicellular organisms and higher species. This allows for the categorization of stress-levels into discrete stages, each with distinct and quantifiable features (for a similar attempt, see Romero et al., 2009). ‘Low’ amounts of stress (prediction error) should be associated with low-level action-perception cycles, i.e. activity of short loops within the nested hierarchy and low levels of permutation entropy in hierarchical message passing ([0] Reflexive, habitual behavior, homeostatic control). This reflects the successful suppression of low-level prediction errors by predictive structures of low model complexity (short stimulus-evaluation-response loops). When stress levels rise to mild or moderate levels, we expect increased involvement of higher level generative models and behavioral policies of corresponding complexity that conspire to suppress rising levels of prediction error. This stage involves increased activity within processing loops of increasing length and rising levels of permutation entropy ([1] goal-directed behavior, allostatic control). In contrast, we expect the activity of higher hierarchical levels to decrease again when stress levels become more severe. This reflects the dissolution of higher-level goal states when the hierarchy is taxed to its limits as a result of hub overload and failure ([2] regression to homeostatic behavior, allostatic overload). Thus, both low [0] and high [2] stress levels should engage habitual rather than goal-directed forms of behavior. The final two stages involve an undercontrolled state of high permutation entropy ([3] critical slowing down, CSD), which predicts a sudden loss of functional or structural integrity ([4] loss of control, tipping points / decompensation). All of these stages can be identified objectively (van de Leemput

et al., 2014; Zhu et al., 2020). Predictions with respect to (the directionality of) network connections and levels of disorder in hierarchical message passing under different levels of stress can be tested using the quantitative measures described above.

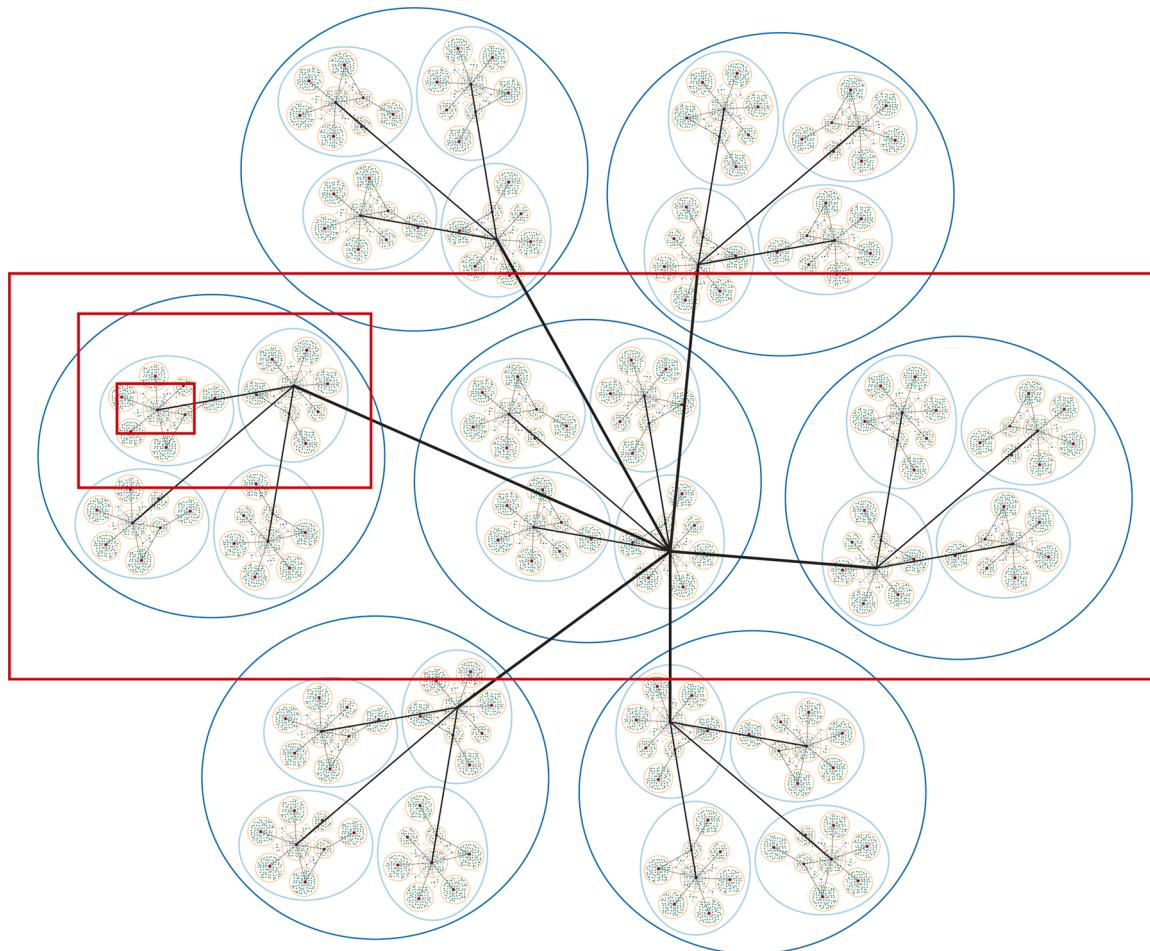
An interesting approach would be to simulate changes in the performance of hierarchical control systems under different levels of stress using artificial systems (e.g. information bottleneck systems, see below). Such studies would allow testing of the hypothesis that folded bow-tie structures with nodes that collapse as a function of node degree as a result of rising levels of prediction error would show a top-down collapse of goal hierarchies (the knots of bow ties), i.e. to produce a model of allostatic overload. This can be done by modeling precision as the weight of connections of prediction error units relative to predictive units. Increases in entropy measures of both hierarchical message passing and behavioral sequences produced by the system should then be predictive of tipping points / loss of homeostasis and serve as a universal model for stagnant growth, disease, or death. Such studies would be a safe way to study the tipping point thresholds as a function of hierarchical depth, providing a generic model for individual differences in fitness. Incidentally, such a model would provide a mechanistic account of the workings of natural selection on organisms that lack adaptive capacity and, thus, link to studies of evolutionary biology.

Finally, it would be interesting to examine to what degree the

structure of human phenotypical networks (inner experience and overt behavior) echoes the physical structure of living network systems as shown in Figs. 6 and 8. Phenotypical networks indeed show signs of small-worldness and nested modular hierarchy (part-whole relationships), as well as statistical dependencies between items that can be explained by physical network architectures capable of hierarchical Bayesian inference (e.g. Goekoop and Goekoop, 2014). A similar approach can be tried in social networks. Here, agent-based simulations could aid in understanding patterns of social interaction at the local level (e.g. mutual dependence or social conflict) as well as global phenomena such as innovation and rumor diffusion, voting, migration, strikes, riot behavior, economic slowdown and warfare.

### 5.3. Modeling organisms: a unified theoretical framework

One of the most interesting features of living systems is that they follow scale-independent rules of network structure and function that apply to all organisms. Such universality means that organisms of any type can be modeled using a minimum set of building blocks under a common theoretical framework. Scholars will not have to make unique models for each organism separately, nor for each level of observation within the organism (e.g. genetic, cellular, systems level, or social). Instead, organisms can be described in terms of a limited set of network



**Fig. 10. Scale-Invariant Features in Organisms Allow for Efficient Modeling.**

*Note:* The scale invariance of biological networks proves useful for modeling organisms. The same network motifs appear at different scale levels of organization, where they support similar functions. For example, red squares indicate the same structural network motif (a folded bow-tie motif) at different spatial scale levels of observation. Modeling organisms would only require knowledge of the number, positions and scale levels of a particular type of motif within an organism, allowing for significant parsimony of description (i.e. organisms can be ‘compressed’ and ‘decompressed’). The fractal-like structure of biological networks means that organisms can be reduced to single feedforward loops at the highest spatial scale level of observation (the level of the individual organism) without losing much information. See text for further details.

motifs (Araujo and Liotta, 2018), allowing for compression of datasets. Additionally, scale invariance means that organisms can be modeled either in all their intricate detail (i.e. the full hierarchy of part-whole relationships) or rather more grossly, as a few global motifs that together perform some global functions, without losing too much information (Fig. 10). Such multilevel ‘coarse graining’ techniques have been shown to be successful in simulating organismic behavior (Derbal, 2013). Here, we show that organisms can be grossly modeled as a giant predictive feed forward loop (FFL; section 3.3), which produces output that provides an update on these predictions via the environment (active inference).

#### 5.4. How biology may inform machine learning

So far, we have discussed how artificial intelligence can help us to understand biological networks in terms of hierarchical Bayesian control systems. Conversely, one may examine how biological systems may inform computer models of hierarchical control systems. For instance, deep networks usually start out with random connections that change after learning. Eventually, the idea of learning is to connect some input (e.g. a series of pixels that together form the shape of a cat) to a desired output (say, the succession of letters (C-A-T) in a non-random fashion by means of a hierarchically organized throughput area that makes these connections. We have seen that such associations are significantly improved when allowing for a hierarchical structure of input, throughput and output modules (Section 2.2). Since non-randomly wired *small world* networks form spontaneously when optimizing the flow of energy through random networks (Jarman et al., 2017), we predict that present-day hierarchical deep networks, when performing at optimal efficiency, must have approached a scale invariant, *small world* network structure. Currently, we know of no studies that have examined existing deep networks directly for small-worldness. A recent study found that fitting a deep network with *small world* network architecture prior to learning significantly enhanced its performance, thanks to the rapid convergence of microstates onto hub states (Javaheripi et al., 2019). A further improvement could be made by fitting deep networks with bottleneck (bow-tie) structure prior to learning (Shwartz-Ziv and Tishby, 2017). Several studies show that information bottlenecks increase the performance of hierarchical (deep) networks by allowing their higher hierarchical levels to perform some kind of compression and generalization of events that take place at lower levels (Hafez-Kolahi and Kasaei, 2019; Shwartz-Ziv and Tishby, 2017). Such performance increases appear to be related to phylogenetic learning (evolution) rather than ontogenetic learning (within-lifespan individual development), hence their introduction may significantly boost system performance by skipping a generic (phylogenetic) learning process, allowing the system to directly proceed with task-relevant (ontogenetic) learning instead. Information bottlenecks may also prove crucial in studies of hierarchical Bayesian inference (interestingly, the objective function used for the free energy principle, i.e. variational free energy, can be cast in terms of compressing and minimum description lengths (Friston, 2019a; MacKay, 1995, 2003; Sun et al., 2011; Wallace and Dowe, 1999). Given the ubiquitous presence of *small world* and bottleneck networks in nature, we expect that such features will soon be detected in hierarchical deep learning systems and that the formation of such structures correlates positively with the performance of such systems. Indeed, the very structure of deep networks necessarily entails a kind of bowtie structure. This is most evident in things like variational autoencoders, which arguably represent the state-of-the-art in deep learning (Zhao et al., 2017). These are deep networks with a bow-tie like architecture that follow the rules of hierarchical Bayesian inference, with a converging input part that is called an ‘encoder’ and a divergent output part that is called a ‘decoder’. Behavior is generated by decoding abstract states into hierarchical output sequences in a top-down manner. We predict that such structures will show biologically plausible behavior when folded to connect input and output structures at corresponding

hierarchical levels (Safron, 2020) and when accounting for hub overload and failure during stress (Stam, 2014). Overall, it is interesting to note that the network architectures that predominate in machine learning (e.g. deep convolution neural networks) conform almost exactly to the principles that we have been exposing, i.e. they have an explicit hierarchical structure with a certain kind of sparsity, following rules of predictive coding and hierarchical Bayesian inference.

As a final remark, biological systems may inspire machine learning techniques with respect to the generic response they show to severe stress and the overtaxing of their hierarchies of control. Lowering integrative control at the cost of contextual integration may be an answer in situations that require rapid decisions within the context of limited energy supply (e.g. battery powered devices). This may speed up system performance in dire situations, e.g. when used in military situations, self driving cars or policing. The prospect of ‘stressed robots’ that weigh selfish and selfless goals may not seem very appealing, but may ultimately prove to be of significant value. For instance, robots may be programmed to never abandon higher level (normative) goals over lower level (self-centered or social) goals in relevant situations, effectively causing them to remain morally just and impartial, or to self-sacrifice (fail for the global good) under stressful conditions.

#### 5.5. Conclusion

To conclude, we have examined how biological network systems have structural features that allow them to function as hierarchical Bayesian control systems. Such systems have generic ways of producing behavior and responding to stress, which may prove useful in understanding animal as well as human behavior. Biology on the other hand keeps on inspiring man-made systems, for which we have made some suggestions. A list of techniques has been presented that can be used to test the hypotheses presented in this paper.

#### Acknowledgements

We thank the reviewers for their constructive comments that added significantly to the quality of the paper.

#### References

- Adams, R.A., Shipp, S., Friston, K., 2013. Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* 218 (3), 611–643.
- Alon, U., 2007. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8 (6), 450–461. <https://doi.org/10.1038/nrg2102>.
- Amodio, D.M., Frith, C.D., 2006. Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* 7 (4), 268–277. <https://doi.org/10.1038/nrn1884>.
- Araujo, R.P., Liotta, L.A., 2018. The topological requirements for robust perfect adaptation in networks of any size. *Nat. Commun.* 9 (1), 1757. <https://doi.org/10.1038/s41467-018-04151-6>.
- Arnsten, A.F.T., 2009. Stress signalling pathways that impair prefrontal cortex structure and function. *Nat. Rev. Neurosci.* 10 (6), 410–422. <https://doi.org/10.1038/nrn2648>.
- Ashby, W.R., 1947. Principles of the self-organizing dynamic system. *J. Gen. Psychol.* 37 (2), 125–128. <https://doi.org/10.1080/00221309.1947.9918144>.
- Ashby, W.R., 1961. *An Introduction to Cybernetics*. Chapman & Hall Ltd.
- Asher, L., Collins, L.M., Ortiz-Pelaez, A., Drewe, J.A., Nicol, C.J., Pfeiffer, D.U., 2009. Recent advances in the analysis of behavioural organization and interpretation as indicators of animal welfare. *J. R. Soc. Interface* 6 (41), 1103–1119. <https://doi.org/10.1098/rsif.2009.0221>.
- Azevedo, F.A., Carvalho, L.R., Grinberg, L.T., Farfel, J.M., Ferretti, R.E., Leite, R.E., Filho, W.J., Lent, R., Herculano-Houzel, S., 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* 513 (5), 532–541. <https://doi.org/10.1002/cne.21974>.
- Badcock, P.B., Friston, K., Ramstead, M., 2019. The hierarchically mechanistic mind: a free-energy formulation of the human psyche. *Phys. Life Rev.* 31, 104–121. <https://doi.org/10.1016/j.plrev.2018.10.002>.
- Bae, M.J., Park, Y.S., 2014. Biological early warning system based on the responses of aquatic organisms to disturbances: a review. *Sci. Total Environ.* 466, 635–649. <https://doi.org/10.1016/j.scitotenv.2013.07.075>.
- Baker, C.L., Jara-Ettinger, J., Saxe, R., Tenenbaum, J.B., 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behavior* 1, 64. <https://doi.org/10.1038/s41562-017-0064>.
- Barabasi, A.L., 2009. Scale-free networks: a decade and beyond. *Science* 325 (5939), 412–413. <https://doi.org/10.1126/science.1173299>.

- Barabasi, A.L., 2013. Network science. *Philos. Trans. Math. Phys. Eng. Sci.* 371, 20120375. <https://doi.org/10.1098/rsta.2012.0375>.
- Barabasi, A.L., Bonabeau, E., 2003. Scale-free networks. *Sci. Am.* 288 (5), 60–69. PM: 12701331.
- Barabasi, A.L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. <https://doi.org/10.1038/nrg1272>.
- Barzel, B., Barabasi, A.L., 2013a. Network link prediction by global silencing of indirect correlations. *Nat. Biotechnol.* 31 (8), 720–725. <https://doi.org/10.1038/nbt.2601>.
- Barzel, B., Barabasi, A.L., 2013b. Universality in network dynamics. *Nat. Phys.* 9, 673–681. <https://doi.org/10.1038/nphys2741>.
- Bassett, D.S., Bullmore, E.D., 2006. Small-world brain networks. *Neuroscientist* 12 (6), 512–523. <https://doi.org/10.1177/1073858406293182>.
- Bekoff, M., Pierce, J., 2009. *Wild Justice: the Moral Lives of Animals*. University of Chicago Press.
- Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112 (518), 859–877. <https://doi.org/10.1080/01621459.2017.1285773>.
- Botvinick, M.M., 2007. Multilevel structure in behavior and in the brain: a model of Fuster's hierarchy. *Philos. Trans. Biol. Sci.* 362 (1485), 1615–1626. <https://doi.org/10.1098/rstb.2007.2056>.
- Botvinick, M.M., 2008. Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci. (Regul. Ed.)* 12 (5), 201–208. <https://doi.org/10.1016/j.tics.2008.02.009>.
- Botvinick, M.M., Weinstein, A., 2014. Model-based hierarchical reinforcement learning and human action control. *Philos. Trans. Biol. Sci.* 369 (1655) <https://doi.org/10.1098/rstb.2013.0480>.
- Boutte, C.C., Croson, S., 2013. Bacterial lifestyle shapes stringent response activation. *Trends Microbiol.* 21 (4), 174–180. <https://doi.org/10.1016/j.tim.2013.01.002>.
- Braatenberg, V., 1984. *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge MA.
- Brembs, B., 2003. Operant conditioning in invertebrates. *Curr. Opin. Neurobiol.* 13 (6), 710–717. <https://doi.org/10.1016/j.conb.2003.10.002>.
- Brooks, R., 1986. A robust layered control system for a mobile robot. *Ieee J. Robot. Autom.* 2 (1), 14–23. <https://doi.org/10.1109/JRA.1986.1087032>.
- Brunner, R., Henze, R., Parzer, P., Kramer, J., Feigl, N., Lutz, K., Essig, M., Resch, F., Stieljes, B., 2010. Reduced prefrontal and orbitofrontal gray matter in female adolescents with borderline personality disorder: Is it disorder specific? *NeuroImage* 49 (1), 114–120. <https://doi.org/10.1016/j.neuroimage.2009.07.070>.
- Buchanan, T.W., Preston, S.D., 2014. Stress leads to prosocial action in immediate need situations. *Front. Behav. Neurosci.* 8, 5. <https://doi.org/10.3389/fnbeh.2014.00005>.
- Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10 (3), 186–198. <https://doi.org/10.1038/nrn2575>.
- Byrd, T.A., Erez, A., Vogel, R.M., Peterson, C., Vennettilli, M., Altan-Bonnet, G., Mugler, A., 2019. Critical slowing down in biochemical networks with feedback. *Phys. Rev. E* 100 (2), 022415. <https://doi.org/10.1103/PhysRevE.100.022415>.
- Calvo, P., Baluška, F., 2015. Conditions for minimal intelligence across Eukaryota: a cognitive science perspective. *Front. Psychol.* 6, 1329. <https://doi.org/10.3389/fpsyg.2015.01329>.
- Calvo, P., Friston, K., 2017. Predicting green: really radical (plant) predictive processing. *J. R. Soc. Interface* 14 (131), 20170096. <https://doi.org/10.1098/rsif.2017.0096>.
- Cannon, W.B., 1929. Organization for physiological homeostasis. *Physiol. Rev.* 9 (3), 399–431. <https://doi.org/10.1152/physrev.1929.9.3.399>.
- Cannon, W.B., 1932. *The Wisdom of the Body*. Norton.
- Cloninger, C.R., 2008. The psychobiological theory of temperament and character: comment on Farmer and Goldberg (2008). *Psychol. Assess.* 20 (3), 292–299. <https://doi.org/10.1037/a0012933> discussion 300–4.
- Commons, M.L., Pekker, A., 2008. Presenting the formal theory of hierarchical complexity. *World Futures* 64 (5–7), 375–382. <https://doi.org/10.1080/02604020802301204>.
- Conant, R.C., Ross Ashby, W., 1970. Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1 (2), 89–97.
- Constant, A., Ramstead, M.J., Veissière, S.P., Campbell, J.O., Friston, K., 2018. A variational approach to niche construction. *J. R. Soc. Interface* 15 (141), 20170685. <https://doi.org/10.1098/rsif.2017.0685>.
- Constant, A., Ramstead, M.J.D., Veissière, S.P.L., Friston, K.J., 2019. Regimes of expectations: an active inference model of social conformity and decision making. *Front. Psychol.* 10, 679. <https://doi.org/10.3389/fpsyg.2019.00679>.
- Corominas-Murtra, B., Goní, J., Solé, R.V., Rodríguez-Caso, C., 2013. On the origins of hierarchy in complex networks. *Proc. Natl. Acad. Sci. U. S. A.* 110 (33), 13316–13321. <https://doi.org/10.1073/pnas.1300832110>.
- Csardi, G., Nepusz, T., 2005. The igraph software package for complex network research. *InterJournal* 1695 (5), 1–9.
- Csete, M., Doyle, J., 2004. Bow-ties, metabolism and disease. *Trends Biotechnol.* 22 (9), 446–450. <https://doi.org/10.1016/j.tibtech.2004.07.007>.
- de Kleijn, R., Kachergis, G., Hommel, B., 2014d. Everyday robotic action: lessons from human action control. *Front. Neurorobot.* 8, 13. <https://doi.org/10.3389/fnbot.2014.00013>.
- de Nadal, E., Ammerer, G., Posas, F., 2011d. Controlling gene expression in response to stress. *Nat. Rev. Genet.* 12 (12), 833–845. <https://doi.org/10.1038/nrg3055>.
- Del Giudice, M., 2020. Rethinking the fast-slow continuum of individual differences. *Evol. Hum. Behav.* <https://doi.org/10.1016/j.evolhumbehav.2020.05.004>.
- Del Giudice, M., Gangestad, S.W., Kaplan, H.S., 2015. Life history theory and evolutionary psychology. In: *The Handbook of Evolutionary Psychology*, Vol. 1: Foundations, pp. 88–114. <https://doi.org/10.1002/9781119125563.evpsych102>.
- Del Giudice, M., Buck, C.L., Chaby, L.E., Gormally, B.M., Taff, C.C., Thawley, C.J., Vitousek, M.N., Wada, H., 2018. What is stress? A systems perspective. *Integr. Comp. Biol.* 58 (6), 1019–1032. <https://doi.org/10.1093/icb/icy114>.
- Derbal, Y., 2013. On modeling of living organisms using hierarchical coarse-graining abstractions of knowledge. *J. Biol. Syst.* 21 (01), 1350008. <https://doi.org/10.1142/S0218339013500083>.
- Dias-Ferreira, E., Sousa, J.C., Melo, I., Morgado, P., Mesquita, A.R., Cerqueira, J.J., Costa, R.M., Sousa, N., 2009. Chronic stress causes frontostriatal reorganization and affects decision-making. *Science* 325 (5940), 621–625. <https://doi.org/10.1126/science.1171203>.
- Doll, B.B., Simon, D.A., Daw, N.D., 2012. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* 22 (6), 1075–1081. <https://doi.org/10.1016/j.conb.2012.08.003>.
- Dunbar, R.I., Shultz, S., 2007. Evolution in the social brain. *Science* 317 (5843), 1344–1347. <https://doi.org/10.1126/science.1145463>.
- Edlund, J.A., Chaumont, N., Hintze, A., Koch, C., Tononi, G., Adamc, C., 2011. Integrated information increases with fitness in the evolution of animats. *PLoS Comput. Biol.* 7 (10), e1002236. <https://doi.org/10.1371/journal.pcbi.1002236>.
- Eguiraun, H., López-de-Ipiña, K., Martínez, I., 2014. Application of entropy and fractal dimension analyses to the pattern recognition of contaminated fish responses in aquaculture. *Entropy* 16 (11), 6133–6151. <https://doi.org/10.3390/e16116133>.
- Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. *Evol. Hum. Behav.* 25 (2), 63–87. [https://doi.org/10.1016/S1090-5138\(04\)00005-4](https://doi.org/10.1016/S1090-5138(04)00005-4).
- Fehr, E., Schurtenberger, I., 2018. Normative foundations of human cooperation. *Nat. Hum. Behav.* 2, 458–468. <https://doi.org/10.1038/s41562-018-0385-5>.
- Feldman, A.G., Levin, M.F., 2009. The equilibrium-point hypothesis—past, present and future. *Progress in Motor Control*. Springer, pp. 699–726. [https://doi.org/10.1007/978-0-387-77064-2\\_38](https://doi.org/10.1007/978-0-387-77064-2_38).
- Fortier, M., Friedman, D.A., 2018. Of woodlice and men: a Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bull.* 2, 17–43. [https://www.aliusresearch.org/uploads/9/1/6/0/91600416/friston\\_-of\\_woodlice\\_and\\_men.pdf](https://www.aliusresearch.org/uploads/9/1/6/0/91600416/friston_-of_woodlice_and_men.pdf).
- Freeman, W.J., 2001. *How Brains Make up Their Minds*. Columbia University Press.
- Freeman, W.J., 2005. A field-theoretic approach to understanding scale-free neocortical dynamics. *Biol. Cybern.* 92 (6), 350–359. <https://doi.org/10.1007/s00422-005-0563-1>.
- Friedlander, T., Mayo, A.E., Slutzky, T., Alon, U., 2015. Evolution of bow-tie architectures in biology. *PLoS Comput. Biol.* 11 (3), e1004055. <https://doi.org/10.1371/journal.pcbi.1004055>.
- Friston, K., 2010. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. <https://doi.org/10.1038/nrn2787>.
- Friston, K., 2012. A free energy principle for biological systems. *Entropy* 14 (11), 2100–2121. <https://doi.org/10.3390/e14112100>.
- Friston, K., 2018. Does predictive coding have a future? *Nat. Neurosci.* 21, 1019–1021. <https://doi.org/10.1038/s41593-018-0200-7>.
- Friston, K., 2019a. A free energy principle for a particular physics. *arXiv:1906.10184 [q-bio.NC]*. <https://arxiv.org/abs/1906.10184>.
- Friston, K., 2019b. Waves of prediction. *PLoS Biol.* 17, e3000426. <https://doi.org/10.1371/journal.pbio.3000426>.
- Friston, K., Frith, C.D., 2015. Active inference, communication and hermeneutics. *Cortex* 68, 129–143. <https://doi.org/10.1016/j.cortex.2015.03.025>.
- Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. *Philos. Trans. Biol. Sci.* 364 (1521), 1211–1221. <https://doi.org/10.1098/rstb.2008.0300>.
- Friston, K., Kilner, J., Harrison, L., 2006. A free energy principle for the brain. *J. Physiol. 100* (1–3), 70–87. <https://doi.org/10.1016/j.jphysparis.2006.10.001>.
- Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., Dolan, R., 2013. The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7 (598). <https://doi.org/10.3389/fnhum.2013.00598>.
- Friston, K., Parr, T., de Vries, B., 2017. The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1 (4), 381–414. [https://doi.org/10.1162/NETN\\_a\\_00018](https://doi.org/10.1162/NETN_a_00018).
- Friston, K.J., Fortier, M., Friedman, D.A., 2018. Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bull.* 2, 17–43.
- Gallagher, H.L., Frith, C.D., 2003. Functional imaging of 'theory of mind'. *Trends Cogn. Sci. (Regul. Ed.)* 7 (2), 77–83. [https://doi.org/10.1016/S1364-6613\(02\)00025-6](https://doi.org/10.1016/S1364-6613(02)00025-6).
- Gallos, L.K., Song, C., Makse, H.A., 2007. A review of fractality and self-similarity in complex networks. *Phys. A Stat. Mech. Appl.* 386 (2), 686–691. <https://doi.org/10.1016/j.physa.2007.07.069>.
- Girvan, M., Newman, M.E., 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 99 (12), 7821–7826. <https://doi.org/10.1073/pnas.122653799>.
- Goekoop, R., Goekoop, J.G., 2014. A network view on psychiatric disorders: network clusters of symptoms as elementary syndromes of psychopathology. *PLoS One* 9 (11), e112734. <https://doi.org/10.1371/journal.pone.0112734>.
- Goelzer, A., Fromion, V., 2017. Resource allocation in living organisms. *Biochem. Soc. Trans.* 45 (4), 945–952. <https://doi.org/10.1042/bst20160436>.
- Goette, L., Bendahan, S., Thoresen, J., Hollis, F., Sandi, C., 2015. Stress pulls us apart: anxiety leads to differences in competitive confidence under stress. *Psychoneuroendocrinology* 54, 115–123. <https://doi.org/10.1016/j.psyneuen.2015.01.019>.
- Gorenstein, E.E., Newman, J.P., 1980. Disinhibitory psychopathology: a new perspective and a model for research. *Psychol. Rev.* 87 (3), 301–315. <https://doi.org/10.1037/0033-295X.87.3.301>.
- Gosak, M., Stožer, A., Marković, R., Dolenšek, J., Marhl, M., Slak Rupnik, M., Perc, M., 2015. The relationship between node degree and dissipation rate in networks of

- diffusively coupled oscillators and its significance for pancreatic beta cells. *Chaos* 25 (7), 073115. <https://doi.org/10.1063/1.4926673>.
- Gosak, M., Marković, R., Dolenšek, J., Rupnik, M.S., Marhl, M., Stožer, A., Perc, M., 2018. Network science of biological systems at different scales: a review. *Phys. Life Rev.* 24, 118–135. <https://doi.org/10.1016/j.plrev.2017.11.003>.
- Gottesman, S., 2019. Trouble is coming: signaling pathways that regulate general stress responses in bacteria. *J. Biol. Chem.* 12, 11685–11700. <https://doi.org/10.1074/jbc.REV119.005593>.
- Griffiths, T.L., Chater, N., Kemp, C., Perfors, A., Tenenbaum, J., Griffiths, T., 2010. Probabilistic models of cognition: exploring the laws of thought. *Trends Cogn. Sci. (Regul. Ed.)* 357–364. <https://doi.org/10.1016/j.tics.2010.05.004>.
- Gross, T., Blasius, B., 2007. Adaptive coevolutionary networks: a review. *J. R. Soc. Interface* 5 (20), 259–271. <https://doi.org/10.1098/rsif.2007.1229>.
- Hafez-Kolahi, H., Kasaei, S., 2019. Information bottleneck and its applications in deep learning. *arXiv:1904.03743 [cs.LG]*.
- Haggard, P., 2017. Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 196–207. <https://doi.org/10.1038/nrn.2017.14>.
- Hausfater, G., Hrdy, S.B., 2017. Infanticide: Comparative and Evolutionary Perspectives. Routledge.
- Hegdé, J., Fellerman, D.J., 2007. Reappraising the functional implications of the primate visual anatomical hierarchy. *Neuroscientist* 13 (5), 416–421. <https://doi.org/10.1177/1073858407305201>.
- Hermanns, E.J., Henckens, M.J., Joels, M., Fernandez, G., 2014. Dynamic adaptation of large-scale brain networks in response to acute stressors. *Trends Neurosci.* 37 (6), 304–314. <https://doi.org/10.1016/j.tins.2014.03.006>.
- Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B., Tomasello, M., 2007. Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *Science* 317 (5843), 1360–1366. <https://doi.org/10.1126/science.1146282>.
- Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., Friston, K., 2019. A multi-scale view of the emergent complexity of life: a free-energy proposal. *Evolution, Development and Complexity*. Springer, pp. 195–227. [https://doi.org/10.1007/978-3-030-00075-2\\_7](https://doi.org/10.1007/978-3-030-00075-2_7).
- Hilgetag, C.C., Goulas, A., 2016. Is the brain really a small-world network? *Brain Struct. Funct.* 221, 2361–2366. <https://doi.org/10.1007/s00429-015-1035-6>.
- Hillebrand, A., Tewarie, P., van Dellen, E., Yu, M., Carbo, E.W.S., Douw, L., Gouw, A.A., van Straaten, E.C.W., Stam, C.J., 2016. Direction of information flow in large-scale resting-state networks is frequency-dependent. *Proc. Natl. Acad. Sci. U. S. A.* 113 (14), 3867–3872. <https://doi.org/10.1073/pnas.1515657113>.
- Hooley, J.M., Gruber, S.A., Scott, L.A., Hiller, J.B., Yurgelun-Todd, D.A., 2005. Activation in dorsolateral prefrontal cortex in response to maternal criticism and praise in recovered depressed and healthy control participants. *Biol. Psychiatry* 57 (7), 809–812. <https://doi.org/10.1016/j.biopsych.2005.01.012>.
- Humphries, M.D., Gurney, K., 2008. Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence. *PLoS One* 3 (4), e0002051. <https://doi.org/10.1371/journal.pone.0002051>.
- Jarman, N., Steur, E., Trengove, C., Tyukin, I.Y., van Leeuwen, C., 2017. Self-organisation of small-world networks by adaptive rewiring in response to graph diffusion. *Sci. Rep.* 7 (1), 13158. <https://doi.org/10.1038/s41598-017-12589-9>.
- Javaheripi, M., Rouhani, B.D., Koushanfar, F., 2019. SWNet: small-world neural networks and rapid convergence. *arXiv:1904.04862 [cs.LG]*. <https://arxiv.org/abs/1904.04862>.
- Jensen, P.A., Zhu, Z., van Opijnen, T., 2017. Antibiotics disrupt coordination between transcriptional and phenotypic stress responses in pathogenic bacteria. *Cell Rep.* 20 (7), 1705–1716. <https://doi.org/10.1016/j.celrep.2017.07.062>.
- Johnson, T., Zhou, S., Cheah, W., Mansell, W., Young, R., Watson, S., 2020. Implementation of a perceptual controller for an inverted pendulum robot. *J. Intell. Robot. Syst.* 1–10.
- Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. *Science* 349 (6245), 255. <https://doi.org/10.1126/science.aaa8415>.
- Kaiser, M., Hilgetag, C., 2010. Optimal hierarchical modular topologies for producing limited sustained activation of neural networks. *Front. Neuroinform.* 4, 8. <https://www.frontiersin.org/article/10.3389/fninf.2010.00008>.
- Kanai, R., Komura, Y., Shipp, S., Friston, K., 2015. Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philos. Trans. Biol. Sci.* 370 (1668), 20140169 <https://doi.org/10.1098/rstb.2014.0169>.
- Karklin, Y., Lewicki, M.S., 2009. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* 457 (7225), 83–86. <https://doi.org/10.1038/nature07481>.
- Kauffman, S.A., 1993. *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press.
- Kauffman, S.A., 1996. *At Home in the Universe: the Search for the Laws of Self-Organization and Complexity*. Oxford University Press.
- Kestring, M.-L., Bredenphol, M., Klenke, J., Westermann, S., Lincoln, T.M., 2013. The impact of social stress on self-esteem and paranoid ideation. *J. Behav. Ther. Exp. Psychiatry* 44 (1), 122–128. <https://doi.org/10.1016/j.jbtep.2012.07.010>.
- Kirchhoff, M., Parr, T., Palacios, E., Friston, K., Kiverstein, J., 2018. The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15 (138), 20170792. <https://doi.org/10.1098/rsif.2017.0792>.
- Kitano, H., 2004. Biological robustness. *Nat. Rev. Genet.* 5 (11), 826–837. <https://doi.org/10.1038/nrg1471>.
- Kitano, H., 2017. Biological complexity and the need for computational approaches. *Philosophy of Systems Biology*. Springer, pp. 169–180. [https://doi.org/10.1007/978-3-319-47000-9\\_16](https://doi.org/10.1007/978-3-319-47000-9_16).
- Koltko-Rivera, M.E., 2006. Rediscovering the later version of Maslow's hierarchy of needs: self-transcendence and opportunities for theory, research, and unification. *Rev. Gen. Psychol.* 10 (4), 302–317. <https://doi.org/10.1037/1089-2680.10.4.302>.
- Koolhaas, J.M., Bartolomucci, A., Buwalda, B., de Boer, S.F., Flügge, G., Korte, S.M., Meerlo, P., Murison, R., Olivier, B., Palanza, P., 2011. Stress revisited: a critical evaluation of the stress concept. *Neurosci. Biobehav. Rev.* 35 (5), 1291–1301. <https://doi.org/10.1016/j.neubiorev.2011.02.003>.
- Kotov, R., Krueger, R.F., Watson, D., Achenbach, T.M., Althoff, R.R., Bagby, R.M., Brown, T.A., Carpenter, W.T., Caspi, A., Clark, L.A., 2017. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* 126 (4), 454. <https://doi.org/10.1037/abn0000258>.
- Krause, J., James, R., Croft, D.P., 2010. Personality in the context of social networks. *Philos. Trans. Biol. Sci.* 365 (1560), 4099–4106. <https://doi.org/10.1098/rstb.2010.0216>.
- Kriegeskorte, N., 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446. <https://doi.org/10.1146/annurev-vision-082114-035447>.
- Lancichinetti, A., Fortunato, S., 2009. Community detection algorithms: a comparative analysis. *Phys. Rev. E* 80 (5), 056117. <https://doi.org/10.1103/PhysRevE.80.056117>.
- Lee, E.-J., Yun, J.H., 2019. Moral incompetency under time constraint. *J. Bus. Res.* 99, 438–445. <https://doi.org/10.1016/j.jbusres.2017.10.043>.
- Li, J., Hua, X., Haubrock, M., Wang, J., Wingender, E., 2012. The architecture of the gene regulatory networks of different tissues. *Bioinformatics* 28 (18), i509–i514. <https://doi.org/10.1093/bioinformatics/bts387>.
- Limanowski, J., Blankenburg, F., 2013. Minimal self-models and the free energy principle. *Front. Hum. Neurosci.* 7, 547. <https://doi.org/10.3389/fnhum.2013.00547>.
- Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciampi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
- Liu, Y., Chon, T.S., Baek, H., Do, Y., Choi, J.H., Chung, Y.D., 2011. Permutation entropy applied to movement behaviors of *Drosophila melanogaster*. *Mod. Phys. Lett. B* 25 (12n13), 1133–1142. <https://doi.org/10.1142/S021798491102684X>.
- Lobier, M., Siebenhühner, F., Palva, S., Palva, J.M., 2014. Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. *NeuroImage* 85, 853–872. <https://doi.org/10.1016/j.neuroimage.2013.08.056>.
- López-Maury, L., Marguerat, S., Bähler, J., 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* 9 (8), 583–593. <https://doi.org/10.1038/nrg2398>.
- Lyon, P., 2015. The cognitive cell: bacterial behavior reconsidered. *Front. Microbiol.* 6, 264. <https://doi.org/10.3389/fmicb.2015.00264>.
- MacKay, D.J.C., 1995. Free energy minimisation algorithm for decoding and cryptanalysis. *Electron. Lett.* 31 (6), 446–447. <https://doi.org/10.1049/el:19950331>.
- MacKay, D.J.C., 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.
- Maistro, D., Donnarumma, F., Pezzulo, G., 2015. Divide et impera: subgoaling reduces the complexity of probabilistic inference and problem solving. *J. R. Soc. Interface* 12 (104), 20141335. <https://doi.org/10.1098/rsif.2014.1335>.
- Mao, Y., Roberts, S., Pagliaro, S., Csikszentmihalyi, M., Bonaiuto, M., 2016. Optimal experience and optimal identity: a multinational study of the associations between flow and social identity. *Front. Psychol.* 7, 67. <https://doi.org/10.3389/fpsyg.2016.00067>.
- Marchetti, G., Barolo, M., Jovanović, L., Zisser, H., Seborg, D.E., 2008. A feedforward-feedback glucose control strategy for type 1 diabetes mellitus. *J. Process Control* 18 (2), 149–162. <https://doi.org/10.1016/j.jprocont.2007.07.008>.
- Maria, G.A., Escós, J., Alados, C.L., 2004. Complexity of behavioural sequences and their relation to stress conditions in chickens (*Gallus gallus domesticus*): a non-invasive technique to evaluate animal welfare. *Appl. Anim. Behav. Sci.* 86 (1-2), 93–104. <https://doi.org/10.1016/j.applanim.2003.11.012>.
- Markov, N.T., Ercsey-Ravasz, M., Van Essen, D.C., Knoblauch, K., Toroczkai, Z., Kennedy, H., 2013. Cortical high-density counterstream architectures. *Science* 342 (6158), 1238406. <https://doi.org/10.1126/science.1238406>.
- Marles-Wright, J., Grant, T., Delumeau, O., Van Duinen, G., Firbank, S.J., Lewis, P.J., Murray, J.W., Newman, J.A., Quin, M.B., Race, P.R., 2008. Molecular architecture of the “stressosome”, a signal integration and transduction hub. *Science* 322 (5898), 92–96. <https://doi.org/10.1126/science.1159572>.
- Mars, R.B., Sallet, J., Neubert, F.-X., Rushworth, M.F., 2013. Connectivity profiles reveal the relationship between brain areas for social cognition in human and monkey temporoparietal cortex. *Proc. Natl. Acad. Sci. U. S. A.* 110 (26), 10806–10811. <https://doi.org/10.1073/pnas.1302956110>.
- Märtens, M., Meier, J., Hillebrand, A., Tewarie, P., Van Mieghem, P., 2017. Brain network clustering with information flow motifs. *Appl. Netw. Sci.* 2 (1), 25. <https://doi.org/10.1007/s41109-017-0046-z>.
- Masoudi-Nejad, A., Schreiber, F., Kashani, Z.R.M., 2012. Building blocks of biological networks: a review on major network motif discovery algorithms. *IET Syst. Biol.* 6 (5), 164–174. <https://doi.org/10.1049/iet-syb.2011.0011>.
- McClelland, J.L., Botvinick, M.M., Noelle, D.C., Plaut, D.C., Rogers, T.T., Seidenberg, M.S., Smith, L.B., 2010. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci. (Regul. Ed.)* 14 (8), 348–356. <https://doi.org/10.1016/j.tics.2010.06.002>.

- McEwen, B.S., Wingfield, J.C., 2003. The concept of allostasis in biology and biomedicine. *Horm. Behav.* 43 (1), 2–15. [https://doi.org/10.1016/S0018-506X\(02\)00024-7](https://doi.org/10.1016/S0018-506X(02)00024-7).
- McEwen, B.S., Bowles, N.P., Gray, J.D., Hill, M.N., Hunter, R.G., Karatsoreos, I.N., Nasca, C., 2015. Mechanisms of stress in the brain. *Nat. Neurosci.* 18 (10), 1353–1363. <https://doi.org/10.1038/nn.4086>.
- McTeague, L.M., Goodkind, M.S., Etkin, A., 2016. Transdiagnostic impairment of cognitive control in mental illness. *J. Psychiatr. Res.* 83, 37–46. <https://doi.org/10.1016/j.jpsychires.2016.08.001>.
- Meeske, A.J., Rodrigues, C.D.A., Brady, J., Lim, H.C., Bernhardt, T.G., Rudner, D.Z., 2016. High-throughput genetic screens identify a large and diverse collection of new sporulation genes in *Bacillus subtilis*. *PLoS Biol.* 14 (1), e1002341 <https://doi.org/10.1371/journal.pbio.1002341>.
- Mendez, M.F., 2009. The neurobiology of moral behavior: review and neuropsychiatric implications. *CNS Spectr.* 14 (11), 608–620. <https://doi.org/10.1017/S109285900023853>.
- Mesulam, M., 1998. From sensation to cognition. *Brain* 121 (6), 1013–1052. <https://doi.org/10.1093/brain/121.6.1013>.
- Mesulam, M., 2008. Representation, inference, and transcendent encoding in neurocognitive networks of the human brain. *Ann. Neurol.* 64 (4), 367–378. <https://doi.org/10.1002/ana.21534>.
- Meunier, D., Lambiotte, R., Fornito, A., Ersche, K.D., Bullmore, E.T., 2009. Hierarchical modularity in human brain functional networks. *Front. Neuroinform.* 3, 37. <https://doi.org/10.3389/neuro.11.037.2009>.
- Meunier, D., Lambiotte, R., Bullmore, E.T., 2010. Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* 4, 200. <https://doi.org/10.3389/fnins.2010.00020>.
- Milgram, S., 1967. The small world problem. *Psychol. Today* 1 (1), 61–67.
- Mitchell, A., Romano, G.H., Groisman, B., Yona, A., Dekel, E., Kupiec, M., Dahan, O., Pilpel, Y., 2009. Adaptive prediction of environmental changes by microorganisms. *Nature* 460, 220–224. <https://doi.org/10.1038/nature08112>.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. *Nature* 518, 529–533. <https://doi.org/10.1038/nature14236>.
- Moore, G.E., Baldwin, T., 1993. *Principia ethica*. Cambridge University Press.
- Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R.J., Friston, K., 2014. Bayesian inferences about the self (and others): a review. *Conscious. Cogn.* 25, 67–76. <https://doi.org/10.1016/j.concog.2014.01.009>.
- Nagabandi, A., Kahn, G., Fearing, R.S., Levine, S., 2018. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. 2018 IEEE International Conference on Robotics and Automation (ICRA). <https://doi.org/10.1109/ICRA.2018.8463189>.
- Nagar, S.D., Aggarwal, B., Joon, S., Bhatnagar, R., Bhatnagar, S., 2016. A network biology approach to decipher stress response in bacteria using *Escherichia coli* as a model. *OMICS* 20 (5), 310–324. <https://doi.org/10.1089/omi.2016.0028>.
- Newman, M.E., 2004. Analysis of weighted networks. *Phys. Rev. E* 70 (5 Pt 2), 056131. <https://doi.org/10.1103/PhysRevE.70.056131>.
- Newman, M.E., 2006. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74 (3 Pt 2), 036104. <https://doi.org/10.1103/PhysRevE.74.036104>.
- Newman, M.E., Girvan, M., 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* 69 (2 Pt 2), 026113. <https://doi.org/10.1103/PhysRevE.69.026113>.
- Newman, M.E., Barabási, A., Watts, D.J., 2006. *The Structure and Dynamics of Networks*. Princeton University Press.
- Northoff, G., Heinzel, A., de Greck, M., Bermudez, F., Dobrowolny, H., Panksepp, J., 2006. Self-referential processing in our brain—a meta-analysis of imaging studies on the self. *NeuroImage* 31 (1), 440–457. <https://doi.org/10.1016/j.neuroimage.2005.12.002>.
- Oltvai, Z.N., Barabasi, A.L., 2002. Systems biology: Life's complexity pyramid. *Science* 295 (5594), 763–764. <https://doi.org/10.1126/science.1078563>.
- Ondobaka, S., Kilner, J., Friston, K., 2017. The role of interoceptive inference in theory of mind. *Brain Cogn.* 112, 64–68. <https://doi.org/10.1016/j.bandc.2015.08.002>.
- Opsahl, T., Colizza, V., Panzarasa, P., Ramasco, J.J., 2008. Prominence and control: the weighted rich-club effect. *Phys. Rev. Lett.* 101 (16), 168702 <https://doi.org/10.1103/PhysRevLett.101.168702>.
- Pascanu, R., Li, Y., Vinyals, O., Heess, N., Buesing, L., Racanière, S., Reichert, D., Weber, T., Wierstra, D., Battaglia, P., 2017. Learning model-based planning from scratch. arXiv:1707.06170 [cs.AI]. <https://arxiv.org/abs/1707.06170>.
- Peters, A., McEwen, B.S., Friston, K., 2017. Uncertainty and stress: why it causes diseases and how it is mastered by the brain. *Prog. Neurobiol.* 156, 164–188. <https://doi.org/10.1016/j.pneurobio.2017.05.004>.
- Pezzulo, G., 2012. An active inference view of cognitive control. *Front. Psychol.* 3, 478. <https://doi.org/10.3389/fpsyg.2012.00478>.
- Pezzulo, G., Rigoli, F., Friston, K., 2015. Active inference, homeostatic regulation and adaptive behavioral control. *Prog. Neurobiol.* 134, 17–35. <https://doi.org/10.1016/j.pneurobio.2015.09.001>.
- Pezzulo, G., Rigoli, F., Friston, K., 2018. Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci. (Regul. Ed.)* 22 (4), 294–306. <https://doi.org/10.1016/j.tics.2018.01.009>.
- Powers, W.T., 1973a. *Behavior: the Control of Perception*. Aldine.
- Powers, W.T., 1973b. Feedback: beyond behaviorism. *Science* 179 (4071), 351–356. <https://doi.org/10.1126/science.179.4071.351>.
- Powers, W.T., Clark, R., McFarland, R., 1960. A general feedback theory of human behavior: part II. *Percept. Mot. Skills* 11 (3), 309–323. <https://doi.org/10.2466/pms.1960.11.3.309>.
- Racanière, S., Weber, T., Reichert, D., Buesing, L., Guez, A., Rezende, D.J., Badia, A.P., Vinyals, O., Heess, N., Li, Y., 2017. Imagination-augmented agents for deep reinforcement learning. *Advances in Neural Information Processing Systems*.
- Ramstead, M.J.D., Badcock, P.B., Friston, K., 2018. Answering Schrödinger's question: a free-energy formulation. *Phys. Life Rev.* 24, 1–16. <https://doi.org/10.1016/j.plrev.2017.09.001>.
- Ravasz, E., Barabasi, A.L., 2003. Hierarchical organization in complex networks. *Phys. Rev. E* 67 (2 Pt 2), 026112. <https://doi.org/10.1103/PhysRevE.67.026112>.
- Ribas-Fernandes, J.J., Solway, A., Diluk, C., McGuire, J.T., Barto, A.G., Niv, Y., Botvinick, M.M., 2011. A neural signature of hierarchical reinforcement learning. *Neuron* 71 (2), 370–379. <https://doi.org/10.1016/j.neuron.2011.05.042>.
- Rohe, T., Noppeneij, U., 2015. Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol.* 13 (2) <https://doi.org/10.1371/journal.pbio.1002073>.
- Rohe, T., Ehli, A.-C., Noppeneij, U., 2019. The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nat. Commun.* 10, 1–17. <https://doi.org/10.1038/s41467-019-09664-2>.
- Romero, L.M., Dickens, M.J., Cyr, N.E., 2009. The reactive scope model—a new model integrating homeostasis, allostasis, and stress. *Horm. Behav.* 55 (3), 375–389. <https://doi.org/10.1016/j.yhbeh.2008.12.009>.
- Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* 52 (3), 1059–1069. <https://doi.org/10.1016/j.neuroimage.2009.10>.
- Ruf, T., Geiser, F., 2015. Daily torpor and hibernation in birds and mammals. *Biol. Rev.* 90 (3), 891–926. <https://doi.org/10.1111/brv.12137>.
- Ruse, M., 1988. *Taking Darwin Seriously. A Naturalistic Approach to Philosophy*. Prometheus.
- Russell Cropanzano, K.J., Citera, M., 1993. A goal hierarchy model of personality, motivation, and leadership. *Res. Organ. Behav.* 15, 267–322.
- Safran, A., 2020. An integrated world modeling theory (IWMT) of consciousness: combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; towards solving the hard problem and characterizing agentic causation. *Front. Artif. Intell.* 3, 30. <https://doi.org/10.3389/frai.2020.00030>.
- Sandi, C., Haller, J., 2015. Stress and the social brain: behavioral effects and neurobiological mechanisms. *Nat. Rev. Neurosci.* 16 (5), 290–304. <https://doi.org/10.1038/nrn3918>.
- Scafetta, N., Marchi, D., West, B.J., 2009. Understanding the complexity of human gait dynamics. *Chaos* 19 (2), 026108. <https://doi.org/10.1063/1.3143035>.
- Scheffer, M., Carpenter, S.R., Lenton, T.M., Bascompte, J., Brock, W., Dakos, V., Pascual, M., 2012. Anticipating critical transitions. *Science* 338 (6105), 344–348. <https://doi.org/10.1126/science.1225244>.
- Scheinies, R., Spirtes, P., Glymour, C., Meek, C., Richardson, T., 1998. The TETRAD project: constraint based aids to causal model specification. *Multivariate Behav. Res.* 33 (1), 65–117. [https://doi.org/10.1207/s15327906mbr3301\\_3](https://doi.org/10.1207/s15327906mbr3301_3).
- Schmidhuber, J., 2015. Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>.
- Schwabe, L., Wolf, O.T., 2009. Stress prompts habit behavior in humans. *J. Neurosci.* 29 (22), 7191–7198. <https://doi.org/10.1523/JNEUROSCI.0979-09.2009>.
- Schwabe, L., Wolf, O.T., 2011. Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. *Behav. Brain Res.* 219 (2), 321–328. <https://doi.org/10.1016/j.bbr.2010.12.038>.
- Seth, A., 2014. The Cybernetic Brain: From Interoceptive Inference to Sensorimotor Contingencies. MINDS Project. MINDS. <https://doi.org/10.15502/9783958570108>.
- Seth, A.K., Metzinger, T., Windt, J.M., 2016. Active interoceptive inference and the emotional brain. *Philos. Trans. Biol. Sci.* 371 (1708), 20160007 <https://doi.org/10.1098/rstb.2016.0007>.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. <https://doi.org/10.1101/gr.123930.3>.
- Shwartz-Ziv, R., Tishby, N., 2017. Opening the black box of deep neural networks via information. arXiv:1703.00810 [cs.LG]. <https://arxiv.org/abs/1703.00810>.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. *Nature* 550, 354. <https://doi.org/10.1038/nature24270>.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV]. <https://arxiv.org/abs/1409.1556>.
- Skinner, B.F., 1990. *The Behavior of Organisms: An Experimental Analysis*. B.F. Skinner Foundation.
- Sleep, C.E., Lyman, D.R., Widiger, T.A., Crowe, M.L., Miller, J.D., 2019. An evaluation of DSM-5 Section III personality disorder Criterion A (impairment) in accounting for psychopathology. *Psychol. Assess.* 31 (10), 1181. <https://doi.org/10.1037/pas0000620>.
- Smith, R., Parr, T., Friston, K., 2019a. Simulating emotions: an active inference model of emotional state inference and emotion concept learning. *Front. Psychol.* 10, 2844. <https://doi.org/10.3389/fpsyg.2019.02844>.
- Smith, R., Schwartenbeck, P., Parr, T., Friston, K., 2019b. An active inference approach to modeling concept learning. bioRxiv, 633677. <https://doi.org/10.1101/633677>.

- Solway, A., Botvinick, M.M., 2012. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychol. Rev.* 119 (1), 120. <https://doi.org/10.1037/a0026435>.
- Solway, A., Diuk, C., Cordova, N., Yee, D., Barto, A.G., Niv, Y., Botvinick, M.M., 2014. Optimal behavioral hierarchy. *PLoS Comput. Biol.* 10 (8), e1003779 <https://doi.org/10.1371/journal.pcbi.1003779>.
- Song, C., Havlin, S., Makse, H.A., 2005. Self-similarity of complex networks. *Nature* 433 (7024), 392–395. <https://doi.org/10.1038/nature03248>.
- Song, C., Havlin, S., Makse, H.A., 2006. Origins of fractality in the growth of complex networks. *Nat. Phys.* 2 (4), 275–281. <https://doi.org/10.1038/nphys266>.
- Sorzano, C.O.S., Vargas, J., Montano, A.P., 2014. A survey of dimensionality reduction techniques. *arXiv:1403.2877 [stat.ML]*. <https://arxiv.org/abs/1403.2877>.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64 (4), 583–639. <https://doi.org/10.1111/1467-9868.00353>.
- Sporrs, O., 2013. Network attributes for segregation and integration in the human brain. *Curr. Opin. Neurobiol.* 23 (2), 162–171. <https://doi.org/10.1016/j.conb.2012.11.015>.
- Sporrs, O., Zwi, J.D., 2004. The small world of the cerebral cortex. *Neuroinformatics* 2 (2), 145–162. <https://doi.org/10.1385/NI:2:2:145>.
- Stam, C.J., 2014. Modern network science of neurological disorders. *Nat. Rev. Neurosci.* 15 (10), 683–695.
- Staniloiu, A., Markowitsch, H., 2012. P-246: the neuroimaging of dissociative disorders. *Eur. Psychiatry* 27, 1. [https://doi.org/10.1016/S0924-9338\(12\)74413-9](https://doi.org/10.1016/S0924-9338(12)74413-9).
- Stankov, L., 2007. The structure among measures of personality, social attitudes, values, and social norms. *J. Individ. Differ.* 28 (4), 240–251. <https://doi.org/10.1027/1614-0001.28.4.240>.
- Starcke, K., Polzer, C., Wolf, O.T., Brand, M., 2011. Does stress alter everyday moral decision-making? *Psychoneuroendocrinology* 36 (2), 210–219. <https://doi.org/10.1016/j.psyneuen.2010.07.010>.
- Storz, G., Hengge, R., 2010. *Bacterial Stress Responses*. American Society for Microbiology Press.
- Sun, X., Zou, Y., Nikiforova, V., Kurths, J., Walther, D., 2010. The complexity of gene expression dynamics revealed by permutation entropy. *BMC Bioinformatics* 11 (1), 607. <https://doi.org/10.1186/1471-2105-11-607>.
- Sun, Y., Gomez, F., Schmidhuber, J., 2011. Planning to be surprised: optimal Bayesian exploration in dynamic environments. *Artificial General Intelligence. AGI 2011. Lecture Notes in Computer Science*. Springer. [https://doi.org/10.1007/978-3-642-22887-2\\_5](https://doi.org/10.1007/978-3-642-22887-2_5).
- Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Upcroft, B., Abbeel, P., Burgard, W., Milford, M., 2018. The limits and potentials of deep learning for robotics. *Int. J. Rob. Res.* 37 (4–5), 405–420. <https://doi.org/10.1177/0278364918770733>.
- Sutton, R.S., Barto, A.G., 2018. *Reinforcement Learning: An Introduction*. MIT Press.
- Tagkopoulos, I., Liu, Y.-C., Tavazoie, S., 2008. Predictive behavior within microbial genetic networks. *Science* 320 (5881), 1313–1317. <https://doi.org/10.1126/science.1154456>.
- Talevich, J.R., Read, S.J., Walsh, D.A., Iyer, R., Chopra, G., 2017. Toward a comprehensive taxonomy of human motives. *PLoS One* 12 (2), e0172279. <https://doi.org/10.1371/journal.pone.0172279>.
- Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: statistics, structure, and abstraction. *Science* 331 (6022), 1279–1285. <https://doi.org/10.1126/science.1192788>.
- Thornton, M.A., Weaverdyck, M.E., Mildner, J.N., Tamir, D.I., 2019. People represent their own mental states more distinctly than those of others. *Nat. Commun.* 10 (1), 1–9. <https://doi.org/10.1038/s41467-019-10083-6>.
- Todd, A.R., Forstmann, M., Burgmier, P., Brooks, A.W., Galinsky, A.D., 2015. Anxious and egocentric: how specific emotions influence perspective taking. *J. Exp. Psychol. Gen.* 144 (2), 374. <https://doi.org/10.1037/xge0000048>.
- Toelch, U., Dolan, R.J., 2015. Informational and normative influences in conformity from a neurocomputational perspective. *Trends Cogn. Sci. (Regul. Ed.)* 19 (10), 579–589. <https://doi.org/10.1016/j.tics.2015.07.007>.
- Tononi, G., Sporns, O., Edelman, G.M., 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. U. S. A.* 91 (11), 5033–5037. <https://doi.org/10.1073/pnas.91.11.5033>.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the 2015 IEEE International Conference on Computer Vision*. <https://doi.org/10.1109/ICCV.2015.510>.
- Ursin, H., Eriksen, H.R., 2010. Cognitive activation theory of stress (CATS). *Neurosci. Biobehav. Rev.* 34 (6), 877–881. <https://doi.org/10.1016/j.neubiorev.2009.03.001>.
- van de Leempt, I.A., Wichers, M., Cramer, A.O., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E.H., Viechtbauer, W., Gilkey, E.J., Aggen, S.H., Derom, C., Jacobs, N., Kendler, K.S., van der Maas, H.L., Neale, M.C., Peeters, F., Thiery, E., Zachar, P., Scheffer, M., 2014. Critical slowing down as early warning for the onset and termination of depression. *Proc. Natl. Acad. Sci. U. S. A.* 111 (1), 87–92. <https://doi.org/10.1073/pnas.1312114110>.
- van den Heuvel, M.P., Stam, C.J., Boersma, M., Hulshoff Pol, H.E., 2008v. Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. *NeuroImage* 43 (3), 528–539. <https://doi.org/10.1016/j.neuroimage.2008.08.010>.
- van den Heuvel, M.P., Kahn, R.S., Goni, J., Sporns, O., 2012v. High-cost, high-capacity backbones for global brain communication. *Proc. Natl. Acad. Sci. U. S. A.* 109, 11372–11377. <https://doi.org/10.1073/pnas.1203593109>.
- van der Meer, L., Costafreda, S., Aleman, A., David, A.S., 2010v. Self-reflection and the brain: a theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. *Neurosci. Biobehav. Rev.* 34 (6), 935–946. <https://doi.org/10.1016/j.neubiorev.2009.12.004>.
- Van Oort, J., Tendolkar, I., Hermans, E., Mulders, P., Beckmann, C., Schene, A., Fernández, G., van Eijndhoven, P., 2017. How the brain connects in response to acute stress: a review at the human brain systems level. *Neurosci. Biobehav. Rev.* 83, 281–297. <https://doi.org/10.1016/j.neubiorev.2017.10.015>.
- Vasil, J., Badcock, P.B., Constant, A., Friston, K., Ramstead, M.J., 2020. A world unto itself: human communication as active inference. *Front. Psychol.* <https://doi.org/10.3389/fpsyg.2020.00417>.
- Veissière, S.P., Constant, A., Ramstead, M.J., Friston, K., Kirmayer, L.J., 2019. Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* 1–97. <https://doi.org/10.1017/S0140525X190001213>.
- Veraart, A.J., Faassen, E.J., Dakos, V., van Nes, E.H., Lürling, M., Scheffer, M., 2012. Recovery rates reflect distance to a tipping point in a living system. *Nature* 481 (7381), 357–359. <https://doi.org/10.1038/nature10723>.
- von Collani, G., Grumm, M., 2009v. On the dimensional structure of personality, ideological beliefs, social attitudes, and personal values. *J. Individ. Differ.* 30 (2), 107–119. <https://doi.org/10.1027/1614-0001.30.2.107>.
- Von Dawans, B., Fischbacher, U., Kirschbaum, C., Fehr, E., Heinrichs, M., 2012. The social dimension of stress reactivity: acute stress increases prosocial behavior in humans. *Psychol. Sci.* 23 (6), 651–660. <https://doi.org/10.1177/0956797611431576>.
- Wallace, C.S., Dowe, D.L., 1999. Minimum message length and Kolmogorov complexity. *Comput. J.* 42 (4), 270–283. <https://doi.org/10.1093/comjnl/42.4.270>.
- Walsh, D.M., 2015. *Organisms, Agency, and Evolution*. Cambridge University Press.
- Watson, E., Yilmaz, L.S., Walhout, A.J., 2015. Understanding metabolic regulation at a systems level: metabolite sensing, mathematical predictions, and model organisms. *Annu. Rev. Genet.* 49, 553–575. <https://doi.org/10.1146/annurev-genet-112414-055257>.
- Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. *Nature* 393 (6684), 440–442. <https://doi.org/10.1038/30918>.
- Wheeloock, M.D., Rangaprakash, D., Harnett, N.G., Wood, K.H., Orem, T.R., Mrug, S., Granger, D.A., Deshpande, G., Knight, D.C., 2018. Psychosocial stress reactivity is associated with decreased whole-brain network efficiency and increased amygdala centrality. *Behav. Neurosci.* 132 (6), 561–572. <https://doi.org/10.1037/bne0000276>.
- Wingfield, J.C., 2003. Control of behavioral strategies for capricious environments. *Anim. Behav.* 807–816. <https://doi.org/10.1006/anbe.2003.2298>.
- Wingfield, J.C., Maney, D.L., Breuner, C.W., Jacobs, J.D., Lynn, S., Ramenofsky, M., Richardson, R.D., 1998. Ecological bases of hormone-behavior interactions: the "emergency life history stage". *Am. Zool.* 38 (1), 191–206. <https://doi.org/10.1093/icb/38.1.191>.
- Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356. <https://doi.org/10.1038/nn.4244>.
- Yan, C., He, Y., 2011. Driving and driven architectures of directed small-world human brain functional networks. *PLoS One* 6 (8), e23460. <https://doi.org/10.1371/journal.pone.0023460>.
- Youssef, F.F., Dookeeram, K., Basdeo, V., Francis, E., Doman, M., Mamed, D., Maloo, S., Degannies, J., Dobo, L., Ditschot, P., 2012. Stress alters personal moral decision making. *Psychoneuroendocrinology* 37 (4), 491–498. <https://doi.org/10.1016/j.psyneuen.2011.07.017>.
- Yu, H., Gerstein, M., 2006. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* 103 (40), 14724–14731. <https://doi.org/10.1073/pnas.0508637103>.
- Zhao, J., Yu, H., Luo, J.-H., Cao, Z.-W., Li, Y.-X., 2006. Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinform.* 7, 386. <https://doi.org/10.1186/1471-2105-7-386>.
- Zhao, T., Zhao, R., Eskcenazi, M., 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv:1703.10960 [cs.CL]*. <https://arxiv.org/abs/1703.10960>.
- Zhu, H., Li, Y., Yuan, M., Ren, Z., Yuan, C., Meng, Y., Wang, J., Deng, W., Qiu, C., Huang, X., Gong, Q., Lui, S., Zhang, W., 2019. Increased functional segregation of brain network associated with symptomatology and sustained attention in chronic post-traumatic stress disorder. *J. Affect. Disord.* 247, 183–191. <https://doi.org/10.1016/j.jad.2019.01.012>.
- Zhu, Z., Surujon, D., Ortiz-Marquez, J.C., Huo, W., Isberg, R.R., Bento, J., van Oprijnen, T., 2020. Entropy of a bacterial stress response is a generalizable predictor for fitness and antibiotic sensitivity. *Nat. Commun.* 11 (1), 1–15. <https://doi.org/10.1038/s41467-020-18134-z>.
- Zimmermann, J., Bohnke, J.R., Eschstruth, R., Mathews, A., Wenzel, K., Leising, D., 2015. The latent structure of personality functioning: investigating criterion a from the alternative model for personality disorders in DSM-5. *J. Abnorm. Psychol.* 124 (3), 532–548. <https://doi.org/10.1037/abn0000059>.
- Zinchenko, O., Arsalidou, M., 2018. Brain responses to social norms: meta-analyses of fMRI studies. *Hum. Brain Mapp.* 39 (2), 955–970. <https://doi.org/10.1002/hbm.23895>.