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

Rutger Goekoop¹ and Roy de Kleijn²

¹Parnassia Group, PsyQ, Department of Mood and Anxiety Disorders, Lijnbaan 4, 2512VA The Hague, Netherlands. Email: R.Goekoop@psyq.nl
²Cognitive Psychology Unit, Leiden University, Wassenaarseweg 52, 2333AK Leiden, Netherlands. Email: rdekleijn@gatech.edu

Abstract
In this paper, we aim to integrate findings from the fields of machine learning and network science to argue that organisms can be modeled as hierarchical Bayesian control systems with small world and bottleneck (bow tie) network structure. The nested hierarchical organization of such networks allows organisms to form increasingly integrated percepts and concepts of their inner and outer context, which can be compared to increasingly encompassing predictive models of the world (goal states), to allow for an optimal control of actions. We argue that hierarchical Bayesian inference produces a hierarchy of goal states, from which it follows that organisms must have some form of ‘highest goals’. For all organisms, these predictive models involve interior (self) models, exterior (social) models and overarching (normative) models. We show how such goals are constructed from progressively lesser goal states and that goal hierarchies tend to decompose in a top-down manner under severe and prolonged levels of stress. This loss of high-level control leads to a disinhibition of subordinate hierarchical levels, producing ‘critical’ behavior and tipping points (a sudden loss of homeostasis). Such phase transitions amount either to disease or the death of the organism. This model can be used to model organisms of any type, including humans. In humans, learning higher-level world models corresponds to personality development. A top-down collapse of high-level integrative goal states under stress is identified as a common factor in all forms of mental disease (psychopathology). The paper concludes by discussing ways of testing these hypotheses empirically.

1 Artificial and biological systems

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 [98]. 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 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 able to reach (super)human level performance in complex tasks without prior instructions [136, 173, 179].
achievement lies in the beginning of the 20th century, when cyberneticists began to model organisms as (perceptual) control systems [6, 33, 34, 156]. 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). 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 to that of the thermostat. The difference between the two (the error) is transferred in some form to the radiator, which tries to close the gap between the desired and actual room temperatures by emitting heat. Many studies have shown that organisms behave in similar ways [25, 156], see Figure 1. Organisms use their input-parts (senses) to monitor the state of their environment and compare their input state to a reference state located within a throughput (evaluation) part. Output (behavior) is then generated until the input matches the setpoint. This helps organisms to find an optimal environmental niche. For example, motor activity in woodlice is varied randomly 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 [198].

Figure 1. 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. 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.

Seminal work by Powers showed that biological systems vary their output freely until the state of the input node I matches a reference value. Their behavior thus serves to keep a percept (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 organisms control their input (percepts) by means of their output (behavior), this type of control is called ‘perceptual control’ [156]. Perceptual control is highly pragmatic: rather than the specific actions, it’s the end-result that counts. By freely ‘emitting behavior’ [181] 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 flexibility and creativity to the production of behavior [157]. 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 [223].

2.2 Organisms as hierarchical control systems
Wiring diagrams such as Figure 1 can only produce rather simple behavior, such as the movement of woodlice. Such diagrams require an extension in order to explain more complex forms of behavior. This includes the formation of specific sequences of actions that allow organisms to accomplish certain tasks. For instance, making a cup of coffee involves a number of simple subtasks (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 cherned coffeebeans, pouring the coffee into a cup, adding milk or sugar, et cetera; [23]). Such output sequences can be more or less efficient depending on the order and the number of recursions in which the subtasks appear [37, 184]. 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 [157, 158, 156]. This hierarchy is symbolized in Figure 1 by the input $r'(t)$ to the setpoint node $T$. This is the output of a higher level control system that is superposed on top of the control system shown in this figure. Figure 2 shows a more elaborate example of a hierarchical perceptual control system.

Figure 2. 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 simple actions (primitives), which we call behavior.
The idea that a hierarchy of (higher order) control systems continuously updates the setpoints of low-level systems (action primitives) is known as the *equilibrium setpoint hypothesis* for motor control [52]. 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; [100]. Rather, organisms produce intermittent bursts of hierarchically organized action sequences that are updated by a repeated sampling of the environment (action–perception cycles) [23]. This is comparable with a musician looking up and down at the scroll sheet every now and then to keep track of the piece. This account of action control remains consistent with perceptual control in the sense that organisms use their (hierarchically organized) action sequences to eventually control their input states. Similar hierarchically organized control systems have been used in e.g. robotics [27].

Since Powers, wiring diagrams of control systems have been modified to explain more 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 [91, 197]). Such systems update their ‘policy’ (input–output strategy) 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 actual reward or value obtained after a previous action (i.e. a memory). These 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 actions would change the input to the system considering previous experiences (*model-based* systems) [46, 185]). Such ‘world models’ are simulations of ideal actions and their possible outcomes (e.g. where different paths in a maze lead to), which are based on 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. Model-based systems require a hierarchical elaboration of their throughput parts, to accommodate complex predictive models of the world. Such goal states are 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 [24].

In the past decade, goal-directed learning has been applied within the context of artificial neural networks [173]. Such networks consist of a layer of input nodes that connect to a layer of throughput nodes (a *hidden layer*), which in turn connect 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 label ‘cat’) with remarkable precision. When deep networks are allowed to construct explicit world models (hierarchical deep belief networks, or deep convolutional systems), their performance increases even further. Such systems can ‘imagine’ a future and formulate articulated goal sequences (goals and subgoals) that are pursued by means of complex action sequences to achieve high success rates [141, 151, 159, 221]. 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 different models of what is going on in the environment and to test these models by performing some kind of action out into the environment. These actions
change the input to the system (via the environment), which serves as a check on model evidence [60]. 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. This ability to ‘change your mind’ has added much to the field of control theory (see below for a further elaboration of active inference).

Thus, adding hierarchy to the output parts of control systems allows for the production of complex action-perception sequences (behavior), whereas adding hierarchy to the throughput (setpoint) parts further boosts the performance of such systems by producing comprehensive goals and subgoals. More recently, studies began to apply hierarchical structure to the input layers of deep networks [136, 180]. The hierarchical structure of perceptive areas had 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) [79, 165]. Hierarchical perception allows control systems to extract increasingly abstract patterns and shapes from raw perceptual input [94, 104, 202]. In 2015, a seminal study was the first to combine hierarchical input (abstract vision) with hierarchical throughput (abstract goal states) and hierarchical output (abstract action control) to produce human-level performance in complex visuospatial tasks (playing Atari computer games) [136]. The system only took raw pixel intensity values as input, after which it autonomously produced a complex series of goals and subgoals (e.g. taking elaborate detours through a maze) and actions 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 [180, 206], medical diagnostics [110] and the game of Go [179] and are quickly finding their way into robotics [199]. In short, 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 organisms as hierarchical control systems 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 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 network structure and function of living systems. This has shown that organisms follow scale invariant rules of network structure and function that apply universally to all living systems (see below). These insights have only partly been integrated with the field of machine learning. The purpose of the current paper is to bring these two influential fields of science further together. We will show that all 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 states and generate increasingly complex action sequences in order to reach (long-term) goals. We then identify a universal response of organisms to contextual cues that overtax the ability of a hierarchy of control systems to correct for environmental disturbances. 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 [134, 217] and the subsequent demonstration that universal laws of network theory govern network structures across a wide range of biological, psychological and social systems [12, 14, 15, 17, 145, 148]. 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 then show that this architecture allows biological systems to function as hierarchical Bayesian control systems and identify a universal response of such structures to extreme forms of stress.

3.1 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 [148]. 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; Figure 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; [68, 144]). 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 (Figure 3).

![Figure 3. Organisms can be conceived of as small world network structures. In such networks, most nodes have few connections whereas some have many (hubs). Hubs connect all nodes in the network in such a way that, on average, each node is a small number of intermediate steps away from any other node in the network (hence the term ‘small world’). In small world networks, hubs contract parts of the network into communities or modules, which are collections of nodes that share more connections amongst themselves than with other nodes. Small world networks are known to produce highly efficient ways 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.

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 [161]). 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 (Figure 4; [13]). The scale-invariance of 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 [67, 187, 186]. Hierarchical network structures such as these form spontaneously under the right conditions (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 [89]. A basic thermodynamic rule therefore suffices to produce modular network structures with short and efficient paths: a phenomenon called self-organization [7, 96]. It has been hypothesized that life kick-started from networks of chemical reactions, which subsequently adapted to meet the more complex demands of life [95, 160].

Figure 4. Schematic representation of scale-invariant structure in small world networks. In biological networks, hub nodes contract sets of other hub nodes into network clusters. Such clusters may themselves cluster into superclusters and so on, producing a hierarchy of part–whole relationships. Each cluster may be represented as a novel node at a higher level of spatial organization (i.e. each node in this figure may represent a cluster). A small world network topology (see text) is found at each spatial scale level of biological organization, which is why this topological feature it is called ‘scale-invariant’ or, ‘scale-free’. Blue node: central hub, which connects a set of 6 red nodes (or clusters) into a single network cluster (or supercluster). 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 (i.e. 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 side view 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; [150])

3.2 The function of biological systems
Biological networks are not just static structures. Energy 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 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’ (i.e. conditional dependencies in time). 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 [16, 58, 72, 99]. 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’ (Figure 5, left image; [42, 99]). The input parts of these structures involve multiple input streams converging onto hub structures, which in turn converge onto higher level hub structures, et cetera, 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 (Figure 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 [15, 216, 226]. A similar organization can be observed in the human brain [119]. Here, a large number of neural columns within the visual cortex (coding for color, texture, speed, orientation, et cetera) 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 hour glass). 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 [9, 30, 54, 130, 131, 132]. 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 structure a scale invariant phenomenon (Box 1) [56, 99, 226].

Box 1: On the structure of organisms: network motifs

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, et cetera). 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 [2, 4, 108]. Studies have found a particular abundance of the so-called feedforward loop (FFL) in living systems [2]. 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 [44]. For instance, the act of eating already increases insulin production regardless of actual increase in blood glucose concentration, which involves a predictive feedforward system [118, 154, 155]. In contrast, feedback motifs contain links that run back from the output nodes to the input nodes, i.e. they are 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 [118]. The activity of feedforward and feedback systems needs to be balanced in order to have the best of both
organisms can be conceived of as nested hierarchal 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. it is a 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 processing to output structures across loops of various path lengths, corresponding to different degrees of information processing (Figure 6). Please note that arrows in this figure only show the global direction of energy flow. Feedforward and feedback connections 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, Figure 6.)

worlds. In biological systems, this balance is what underlies hierarchical predictive coding, which requires bow tie network architecture (see next section, Figure 6). Here, 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. Thus, feedback and feedforward loops run up and down the bow tie hierarchy to balance prediction with prediction errors. 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 [26, 32, 58, 82, 135, 200]. In short, the nested hierarchical bottleneck structure of bow tie motifs and their constituent submotifs provide an optimal infrastructure for hierarchical Bayesian inference. Since bow tie organization is a scale-invariant feature [99, 226], living systems are wired to function as hierarchical Bayesian control systems at every scale level of their anatomy [58, 82, 160]. See Box 2 and next section for further information.

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 [99]. Simulation studies show that hierarchical networks spontaneously evolve bow tie structure under some restrictions [56]. 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 [56]. 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 [188]). Several studies show that information bottlenecks increase the performance of hierarchical (deep) networks by allowing for some form of compression and generalization of events that take place at lower levels [76, 178]. Apparently, a good representation is a simple representation that minimizes complexity cost and uses the fewest degrees of freedom to model the environment, i.e. Occam’s principle [116]. Organisms can therefore be conceived of as dimension reduction machines that perform a clustering on input in an attempt to find the most parsimonious representation without losing too much information.

In recent years, insights have grown that a nested hierarchical structure does not only allow organisms to produce more parsimonious descriptions of events that take place at lower levels of the hierarchy. Rather, higher hierarchical levels appear to represent the hidden causes of the effects (events, activity) that take place at lower levels of organization. This process is called hierarchical Bayesian inference [202]. In this view, each level within the hierarchy generates a model of the hidden causes that produce the effects observed at a lower level. 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 [21, 202]. This property of higher models has been linked to the concept of emergence in complex systems [74, 123]. The idea that biological organisms engage in an active form of hierarchical Bayesian inference has produced an explosion of literature in the past decade (Box 2). So far, however, it has remained largely unclear how this type of processing is implemented in biological systems. To our knowledge, this is the first time that concepts from network science and graph theory (i.e. nested modular bow tie structures) are put together with message passing and probabilistic graphical models in hierarchical Bayesian inference, e.g. [64]. Because of their peculiar network architecture, organisms can combine hierarchical perceptual inference (input) with hierarchical inference of goal states (throughput) and hierarchical action control (output), allowing them to function as hierarchical Bayesian control systems. The flow of information across such structures is not a simple process with energy flowing 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 (Figure 5, right image). This produces short input–output loops near the bottom of the hierarchy as well as longer loops that run from input to output along progressively longer loops, reflecting different degrees of processing (Figure 6). Additionally, feedback loops and feedforward loops run up and down the hierarchy, reflecting predictive coding as well as corrections of such predictions by means of novel input (Box 1, Figure 6). Such structures differ from hierarchical control systems that are traditionally used in engineering and computer science and come with specific functionality (Figure 6). This will be discussed in the next section.

**Box 2: On the function of organisms: the active inference principle**

According to active inference theory, the dynamics of biological systems follows from the basic laws of thermodynamics, i.e. organisms 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 reacting to their input [65]. 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 [58]. 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 a hierarchical series of predictive models (goal states and subgoals) and by testing these models against incoming input by performing actions (‘active inference’). In other words, organisms act to maximize sensory evidence for their own predictions: they are ‘self-fulfilling prophecies’. 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, including microbes and social systems [160]. According to the free energy principle, organisms ‘are’ embodied (Bayesian) models of the world and natural selection is nature’s way of performing Bayesian model selection [82].

4 Organisms: bow tie network structures that function as hierarchical Bayesian control systems

4.1 Input, throughput and output
Intuitively, it is easy to see how a nested modular (hierarchical) bow tie network structure may help organisms to function as hierarchical Bayesian control systems. Organisms must learn to 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 [75]. Any failure to connect stimuli with adaptive responses during the course of their lives (ontogenetic learning) will cause the organism to be eliminated through natural selection (phylogenetic learning). Input–output connections are made by throughput parts, which compare actual input to a predictive model in order to produce an adaptive response. Connecting a suitable output to a particular set of input signals 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. \(i(\text{food} | \text{predator})\), instead of \(i(\text{food}), i(\text{predator})\)). Because of their peculiar structure, nested modular network structures are optimally suited to produce such integrated representations [83]. The input parts of these structures involve multiple inputs that converge onto fewer hub structures (Figure 5). Like spiders in a web, such input 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 substrates (e.g. \(i(\text{food} | \text{predator})\)). Such functional integration goes on until the top of the nested hierarchy of network clusters has been reached. At each level of integration within the input hierarchy, input states are compared to goal states at a similar level of integration and abstraction (e.g. \(t(\text{food} | \text{predator})\)), as well as to a higher level (e.g. \(t'(\text{food} | \text{predator} | \text{hungry})\)) after which the ensuing predictions are disseminated by divergent connections across a number of output hubs to produce an adaptive response. These output hubs in turn impose their integrative states onto a larger number of subordinate output hubs and so on, enabling output sequences (see section 1). Thus, functional integration within the nested modular network structures is an important ingredient for hierarchical perception (input), goal setting (throughput) and action control (output).

Although functional integration in nested modular network structures is a necessary precondi-
tion, it is not sufficient for hierarchical Bayesian inference. The wiring pattern of Figure 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 [1, 59, 61, 64, 93]. 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 (Figure 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 updated ‘beliefs’ (‘empirical priors’). Thus, higher level models attempt to suppress (‘explain away’) prediction errors produced by lower-level systems, whereas lower-level systems in turn try to correct higher-level predictions. Such circularly causal relationships produce oscillations that are typically observed in neural dynamics. The output hierarchy shows a similar but inverted makeup [1, 61]. 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). Incidentally, this means that output signals involve predictions rather than ‘commands’ [1]. Thus, each level within the input hierarchy tries to explain away prediction errors produced at lower levels within the hierarchy by means of (predictive) connections (Figure 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 integrated) model (see section 4.3). 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’). It is thought that similar principles of hierarchical inference apply in any organism, from microbe to man [58, 59, 82, 160]. 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 [58]. In Figure 6, we show the putative wiring scheme for hierarchical predictive coding in biological systems [1, 59, 93], which has been adapted to accommodate an information bottleneck (‘bow tie’ or ‘hour glass’) structure. As can be seen in this figure, bottleneck motifs can be observed at every scale level of organization (i.e. it is a scale invariant feature): ascending and convergent input, descending and divergent output, as well as bottleneck ‘throughput’ structures (predictive or prediction error units) can be observed at the level of individual nodes, clusters and the network structure at large. Indeed, the organism itself can be modeled as one giant feedforward loop (FFL, Box 1), which produces predictive output that feeds back into the organism through the environment (‘active inference’, Box 2, Figure 6B).

At this point, it is important to emphasize the difference between traditional notions of hierarchical inference in statistics and hierarchical inference as it takes place in living systems. First, statistical models usually involve a single hierarchy of generative models. In living systems, the hierarchical architecture of generative models acquires two streams: a sensory or input stream that is primarily concerned with inferring “what the world is doing” and an executive or output stream that infers “what the organism is doing” (either in terms of motor behavior or autonomic function): the inverted hierarchy in Figure 6. Within input hierarchies, increasingly comprehensive perceptual models are produced that try to explain lower-level sensory events (hierarchical perceptual inference) [63]. Within output hierarchies, hierarchical inference corresponds to organisms making increasingly complex predictions with respect to the causes of sensory events in motor (e.g. pro-
Figure 6. Putative wiring scheme that supports hierarchical Bayesian inference in nested modular network structures with information bottleneck (bow tie) organization. Adapted from Adams, Shipp, and Friston [1], Friston [59], and Kanai et al. [93]. Subfigure (a): Excerpt of Figure 5, showing only a single input and output hierarchy for visualization purposes. Each additional level within the nested hierarchy represents an extended generative model at a higher level of integration. (Modules of) black nodes encode predictive states (setpoints, updated predictions, goal states). (Modules of) red nodes encode prediction errors. Blue connections: inhibitory predictions. Red connections: excitatory prediction errors. Connections that give rise to the nested hierarchy of network clusters have been largely omitted (gray connections) for visualization purposes. Input hierarchy: predictive hub nodes 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 through convergent and excitatory connections, reflecting the bottom-up correction of higher level predictions (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 to the low-level (predictive) setpoints of action primitives to produce a concerted response, making output theory-driven and predictive rather than reactive [1]. Note that hub nodes of prediction error clusters at some level within the input hierarchy act as predictive nodes (empirical priors) at the next level of organization, whereas hub nodes of predictive clusters 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. 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 [59, 93]. Subfigure (b): More abstract wiring diagram based on connections shown in subfigure (a), with one more level added when compared to subfigure (a), adapted from Kanai et al. [93]. The nested bottleneck (bow tie) structure is reflected by the copy number of nodes (or clusters), which decreases when ascending in the hierarchy. Note that horizontal cross-connections (gray) allow energy to travel across loops of different lengths. Short stimulus-response loops corresponds to simple (and more complex) reflexes and instinctual responses, whereas progressively longer loops enable habitual, instrumental and goal-directed behavior. See text for further details. Context/environment: 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.

Hierarchical control (e.g. endocrine or enterceptive) areas, which pose as complex behavioral strategies (hierarchical action control) [1]. Such strategies may eventually reduce prediction error within the input hierarchy by changing the environment and, subsequently, the input to the system (‘active inference’). This dual aspect of hierarchical inference is emphasized by referring to hierarchical nested bow tie network architectures (with small world characteristics). This means that "bow tie" should be read as a dual-aspect hierarchy responsible for making inferences about hidden states of the world and action 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 [215], yet seems to be clearly definable from the perspective
of organisms as hierarchical control systems. As observed in section 1, 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 [157]. 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 [185]. 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 Figure 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. As a result, goal states can be construed as posterior expectations and beliefs about controllable but hidden states of the world. Instead of being fixed and given, such goal states are progressively constructed within the narrowing bottlenecks of bow tie structures, which lie in between input and output hierarchies. Thus, living systems acquire intermediate throughput streams that try to infer “what the organism needs to be doing”. Such streams encode a nested hierarchy of goal states [155], which allows organisms to reach a general goal through a series of subgoals and corresponding output strategies. Thus, 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).

In hierarchical inference, organisms need to decide whether signals come from a common cause and should be integrated, or whether they come from independent sources and should be segregated [164]. The number of independent clusters in a nested modular hierarchy is therefore a function of the number of independent contextual cues that are relevant to the organism. When their environments grow more complex, organisms need to expand their hierarchical trees in order to encode more articulated models [49]. For instance, some organism can be confronted with food (A), a mate (B), a rival (C) and a predator (D), all at the same time. It then has to choose whether to eat (1), mate (2), fight (3) or hide (4), given its own internal state hungry (a), alone (b), wounded (c), or weak (d). More complex environments therefore require organisms to increase the number of network clusters within input areas to allow for percepts with higher granularity (‘functional segregation’ or ‘factorization’, i.e. A, B, C, D, a, b, c, d). The functional integration of such clusters into a hierarchy of part-whole relationships allows for deeper forms of perceptual inference, producing more complex percepts (e.g. A/B/C/D/a/b/c/d). Similarly, adding clusters to the output parts allows for a richer repertoire of actions (1, 2, 3, 4) and a deeper integration across these output clusters allows for more elaborate forms of action control and behavior (e.g. ‘courtship’, which may involve complex action sequences of e.g. 3|1|4|2, et cetera). However, having a richer repertoire of actions raises chances that output options will conflict with one another, given a certain input. In living systems, such conflicts are resolved in a hierarchical fashion within the throughput parts of their networks (e.g. t(A|a) \rightarrow 1, t(A|B|b) \rightarrow 2, t(A|B|C|a|b) \rightarrow 3, t(A|B|C|D|a|b|c|d) \rightarrow 4). More complex environments therefore require organisms to expand their goal hierarchies in order to connect input patterns to suitable output patterns for all possible combinations of events (i.e. policy selection). This produces extensive hierarchies of goal states, or ‘deep goal hierarchies’ [155, 154], which consists of progressively longer loops that run from input to output via different levels of integration within the goal hierarchy (Figure 6). During the course of their lives, organisms construct their goal hierarchies (world models) through a process of associative learning, which involves the selective strengthening and weakening of connections within these throughput loops. The rewiring of different parts of the goal hierarchy have been linked to different forms of associative learning [155, 154]: short stimulus-response loops represent simple autonomous and/or motor arc reflexes that allow for basic Pavlovian (stimulus–stimulus) learning and instinctive behavior, respectively. Longer loops allow for more complex forms of behavior such as habit learning and instrumental
learning, whereas the longest loops involve the formation of explicit world models that inspire full goal-directed behavior. These various forms of associative learning are thought to be universal to organisms at any scale level, with increasingly more integrated forms of goal-directed learning occurring within increasingly 'higher' organisms. Pavlovian learning has been observed to occur within bacteria [31], whereas goal-directed learning is observed within higher vertebrates and some invertebrates [153]. Below, we will examine which types of world models are produced within deep goal hierarchies and discuss their putative positions within a nested hierarchy of network clusters.

4.2 A taxonomy of goal states

The central tenet of hierarchical Bayesian inference in biological networks is that organisms aim 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 [58]. From both observational data and theoretical considerations, organisms are thought to construct at least two global types of predictive world models (goal states) at the top of a goal hierarchy. One of these involves a model of the organism itself [109, 138]. 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. Similarly, organisms can infer the (hidden) common causes of effects (input) produced by factors outside of their own body [11, 109, 149]. 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.

As observed in section 4.1, the complexity of a regulatory 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 functional integration versus segregation of network communities 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; [62]. These hypotheses can be tested by comparing the hierarchical network structure of solitary and social species, or social species that differ in their level of sociability (see Discussion).
The same goes for the integration of contextual cues across time. Predictive goal states may involve the current state of the organism and/or its environment but also its future input states. For instance, my current input state $i_t'(A|B|C)$ (= being warm, well fed, no predators) seems to match my current goal state $t_t'(A|B|C)$ and output pattern $o_t'$ (lying down), but this policy may well conflict with my future input state (predictive model) $t'_{t+1}(D|E|F)$ (e.g. being cold, hungry, lurking predators), and corresponding output $o'_{t+1}$ (heating, eating, locomotion; [100]). To resolve such temporal conflicts, the same principle of hierarchical control that allows organisms to integrate increasing numbers of contextual cues in (interpersonal) space can be used to integrate contextual cues across time: temporally more distant goal states can be modeled by control systems that are superposed onto those that predict temporally more proximal ones in a hierarchy of part-whole relationships [155]. This can be seen as a further extension of the hierarchies of interior (self) models and exterior (social) models to accommodate long-term predictions with respect to self and others. Prediction errors that are produced relative to such higher level goal states may result in actions at a time when such events have not yet taken place. Such anticipatory actions may be aimed at a future version of the individual itself or some other agent (e.g. offspring), rather than the current self or the current other (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. Thus, we predict that temporally more distant goal states require more extensive hierarchies of control, i.e. the integration across a larger number of network communities. This hypothesis can be tested by comparing the hierarchical network (bow tie) structure of individuals or species that differ in their ability to anticipate events.

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) representation 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 [202]. 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 (Figure 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 [39, 204]. Thus, hierarchical Bayesian inference predicts that, eventually, goal states are produced that (according to the organism) have a 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 social groups, 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. Conversely, such goals may cause some members of a group to punish themselves or others for social norm violation [51]. Altruistic and law-abiding forms of behavior have been observed in a variety of (higher) organisms [20]. 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’. In the active inference literature, such goals are prior beliefs that rest upon each
organism’s place in a particular eco-niche (with 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 [38, 211]. In humans, the exchange of such insights may aid in an attempt to eventually construct a globally held world view. Thus, normative goals give rise to our scientific, moral and judicial institutions.

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 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, et cetera). Combinations of such activities are typically (but not exclusively) found in so called ‘higher organisms’, and should link to measures of nested hierarchical trees in biological networks (see Discussion). Incidentally, 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 extensions may be 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 contextual comprehensiveness [24, 58].

4.3 Stress in hierarchical Bayesian control systems

The hierarchical control systems perspective on living organisms allows for a clear definition of ‘stress’ [152]. Stress can be defined as the overall level of prediction error, namely, the differences between empirical priors and posterior expectations. In predictive coding formulations (of hierarchical Bayesian control) the overall amplitude of prediction errors is often quantified in terms of precision. 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. This will become important below, when we think about stress as reducing prior precision and rendering the organism more exposed to belief updating based upon immediate...
sensory information. Prediction errors produce an output sequence or stress response to reduce the error. In hierarchical Bayesian control systems, 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; [171]). If error signals cannot be sufficiently countered by a simple, straightforward response generated at the same level of the hierarchy (e.g. side-stepping), prediction error is ‘escalated upward’ to higher level control systems that involve more comprehensive world models and corresponding policies (e.g. walking around the object; [100]). Thus, prediction errors pass a hierarchical succession of goal states (increasingly integrated world models) in an attempt to solve the problem. This corresponds to the application of increasingly complex generative models and corresponding behavioral strategies to suppress prediction errors. Although more elaborate, such strategies may eventually suppress prediction errors in ways that simpler loops cannot (e.g. by reaching the top of a fruit tree, or climbing a social hierarchy).

Since any hierarchy of control systems is finite, prediction error signals (variational free energy) 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 behavioral strategies are ineffective at suppressing prediction errors). This accumulation of prediction error can be thought of in terms of a loss of control over free energy (increased uncertainty, or decreased precision). In the Bayesian inference literature, rising levels of free energy are usually accompanied by increases in computational complexity (i.e. a loss of the ability to compress, see above). This is universally accompanied by a loss of metabolic or thermodynamic efficiency. Thus, the accumulation of prediction error within the higher levels of the hierarchy corresponds to increased energy demand. Indeed, scholars have linked the amount of regulatory activity of organisms to the expenditure of energy and resources [69, 125]. Short loops in a nested hierarchy produce relatively simple (reflexive, instinctive, or habitual) behavior across smaller timescales that require small amounts of regulatory work and energy expenditure (automatic processing, e.g. walking). Longer loops produce more complex (instrumental and goal-directed) behavior across larger timescales that involves more regulatory work and demands more energy (effortful processing, e.g. climbing). Since high-level (effortful) strategies demand more energy, the fruitless pursuit of such strategies may at some point cause energy demand to exceed energy supply. In such cases (allostatic overload; [125]), organisms are known to fall back from higher level to lower forms of behavior to save energy and resources [69, 174]. When energy shortage is particularly severe, organisms revert to a universal stress response, called the emergency (life history stage) response in higher organisms or the stringent response in microbes [43, 220]. Such responses are highly conserved across species, e.g. with *E. coli* bacteria shutting down many contextually redundant regulatory genetic pathways when starved under low-glucose conditions [48], *B. subtilis* entering a spore stage [128] and vertebrates entering some form of energy conservation or hibernation under unfavorable conditions [168]. Emergency responses universally involve a down-regulation of metabolism, reduced motility, reduced sociability, a halt on reproductive activity and/or a tendency to neglect (or even eat) offspring [78, 219]). Overall, they economize on long-term, reproductive and/or altruistic activities in favor of short-term, self-repairing and self-preserving activities. In other words, organisms appear to abandon contextually more integrated (social, normative) goal states in favor of simpler (self-referential) goals. We therefore propose that allostatic overload interferes with the ability of organisms to maintain highly integrated and energy-demanding goal states under severe levels of stress, forcing them to ‘downgrade’ to less integrated goal states that require less regulatory work. 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 emergency responses in multicellular organisms or stringent responses in bacteria. Such responses are mounted in a similar way across species, regardless of the stressor the organism encounters.
(e.g. starvation, drought, heat, cold, acidity, salinity, DNA damage, social stress) [140, 113, 112, 120, 142, 195], or even whether the response is genetic or neuronal in nature [80, 112, 124, 175, 174, 209]. In all cases, vertically accumulated prediction error (free energy) is sensed by a small number of hub structures located near the top of the hierarchy. These hubs represent the highest level of integration across subordinate network clusters (Figure 5, 6, 7) and subsequently have the highest degree or number of connections per node in the network. 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. When energy demand exceeds energy supply, these hub structures simply shut down [73]. This process may be facilitated by glucocorticoids or analog signaling substances that actively degrade the connections that encode higher-order predictive models, as part of the revision of maladaptive models and corresponding policies [152]. Thus, when consistently flooded by ascending prediction error signals, high-order hub structures will drop out of the network. Since these hubs maintain many long-distance connections with other network clusters and subclusters, their shutdown causes the network to fall apart (Figure 8). This causes a shift in the balance between functional integration and segregation of network clusters in favor of functional segregation [205], corresponding to a collapse of the nested modular hierarchy: more encompassing goal states effectively ‘decompose’ into their constituent (sub)components, inducing a corresponding change in behavior. We predict that this collapse is a function of node degree: the most integrative goal states are the first to go, but subordinate levels with lesser-degree hubs 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, et cetera), allostatic overload may dissolve a hierarchy of part–whole relationships, leaving only the local troops and the odd sergeant major to take care of the problem (Figure 8). This hypothesis can be tested by examining measures of hierarchical complexity of network structures at different levels of stress (ascending prediction error, see Discussion).

As discussed, high-level nodes and clusters integrate information across a large number of subordinate models (functional integration). When such hub structures shut down, the individual network clusters that constitute higher level generative models will no longer be connected and start to operate on their own (functional segregation). In other words, the top of the regulatory pyramid is lost (gray nodes in Figure 8). This corresponds to a loss of high-level goal states (precise empirical priors) that try to explain away prediction errors, producing a persistent failure to model and predict high-level structure in the sensorium (higher levels can no longer provide precise or confident constraints on lower levels of inference). From a physiological point of view, this would look like a loss of anticipatory long-term goal-directed behavior (allostasis) and a reversion to short-term homeostasis. Thus, stress causes organisms to take lesser amounts of contextual information into account when formulating a response. Such decontextualization has several short-term advantages. First, less energy and resources are spent on processing stimuli, allowing organisms to endure unfavorable conditions for longer periods of time. Second, the bypassing of higher-level control systems reduces the path length of the network, allowing signals to travel from input to output structures across shorter distances, producing faster responses. Finally, the top-down collapse of high level integrative control amounts to a loss of top-down inhibitory (predictive) connections that constrain lower-level systems. This allows lower-level systems to be expressed more freely, making them more pronounced and easier to trigger, or what is referred to as disinhibition in the psychological sciences [71]. This heightening of the senses (within the input hierarchy) and livening of responses (within the output hierarchy) may provide organisms with just the edge needed to suppress prediction errors, by forcing themselves a way out of a dire situation.

Such changes come at a price, however, which is a loss of regulatory finesse. A decrease of higher level integrative control makes organisms more vulnerable to environmental conditions that require
Hierarchical Bayesian control systems allow organisms to incorporate an increasing number of contextual cues from their environment and create a hierarchy of world models or goal states that are used to inform behavior. At the highest levels of integration, these involve internal (self) models, external (social) models and cross-cutting (normative) models (Figure 7). In severely stressed organisms, the hierarchy of goal states collapses in a top-down manner to save energy and resources. This amounts to a pruning of the nested modular (hierarchical) tree (panel A). We propose that this high-level shut-down involves a shift in the balance between functional integration and segregation of network clusters in biological networks in favor of functional segregation. This process causes a graded decomposition of high-level integrative goal states that are not immediately relevant to the current situation (gray nodes in this figure). This results in a 'decontextualization' of behavior, causing organisms to favor short-term and self-centered actions (informed by self models) over long-term social and/or normative behavior (social and normative models). The regulatory collapse may involve several hierarchical levels of integration, depending on the error levels that are encountered. Hence, rising levels of prediction error (stress) will cause energy to flow from input to output via increasingly short processing loops (blue arrow in panel B). Phenotypically, these changes manifest as organisms ‘downgrading’ from goal-directed to instrumental, habitual or even reflexive forms of behavior. The top-down loss of integrative control produces a disinhibition of lower levels, increasing the risk of critical slowing down. Note that individual differences may cause some organisms to retain top-down control under severe levels of stress. See text for further details.

Additionally, a deep collapse of a regulatory hierarchy may lead to a situation where input almost immediately produces output and vice versa. In such a case, even a small environmental disturbance may trigger an intense, volatile, and uncoordinated response from which it may recover only slowly. In other words, a deep collapse increases the reactivity of a system, making it more strongly coupled to its environment. When the loss of control is particularly severe, the system reaches a state in which the connectivity between the...
different elements of the system becomes so strong that only little energy is needed to trigger a massive, synchronized response: a sudden phase transition, or tipping point. This state is known as ‘critical slowing down’ (CSD; [10]). CSD is a highly generic state of network systems that are poised on the brink of a tipping point. The dynamics of such systems is characterized by high levels of variance, increased anticorrelation between different elements of the system and slow recovery from perturbation [107]. Although CSD is well-known phenomenon in abiotic systems, it has only recently been demonstrated in living systems, (e.g. [212], including human mood fluctuations [107]). In control systems, CSD corresponds to a state of intrinsic hyperconnectivity between the input, throughput and output parts of the system, to such a degree that it increases its extrinsic connectivity (i.e. the system with its environment). 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 of the system will linearly follow that of the environment. This corresponds to a loss of homeostasis, which amounts either to disease or the death of the organism.

To our knowledge, this is first time that a (graded) collapse of higher level control under stress is considered as the primary factor that triggers hyperconnectivity (CSD) and tipping points in living systems. These states may develop when such a collapse reaches too deeply down a hierarchy of control systems (i.e. when the hierarchical tree is pruned beyond a level of adequate control). According to this model, organisms may differ in their susceptibility to criticality and tipping points as a result of individual differences in the development and maturation of their regulatory hierarchies: different heights of the regulatory tree produce different sensitivities to CSD and tipping points (decompensation). Such differences may become particularly manifest when poorly developed hierarchical trees are pruned under stressful conditions beyond a level of adequate control. This hypothesis may be tested e.g. by examining the degree to which measures of hierarchical complexity predict the threshold at which different organisms (e.g. microbes) collapse under extreme conditions (see Discussion [142]).

Some scholars define stress specifically as a failure of control [44], which would situate that state somewhere in between criticality and tipping points as defined above. Others focus more on physiological states [125] or cognitive processes [102, 207]. Here, we employ a more liberal definition of stress as the (cumulative) error state of hierarchical control systems [152]. The advantage of this definition is that it lies on a continuum, with clear and objectifiable stress-responses marking discrete levels of stress, i.e. (0) Routine performance (low-level action-perception cycles, reflexive, instinctive or habitual behavior, ‘effortless control’), (1) Creative problem solving (upward escalation of prediction error signals, instrumental or goal-directed action, ‘effortful control’), (2) Emergency responses (shut down of hub structures, decomposition of non-vital goal states, downgrading from higher to lower forms of associative learning), (3) Critical slowing down (hyperconnectivity) and (4) Tipping points (decompensation). For a similar categorization of the stress response, see [166].

4.4 Example: the human brain as a scale-invariant, 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 [30, 84]. Its 86 billion neurons [8] form neural modules that are an average of around 5 degrees of separation apart from any other module in the brain [18, 85, 191, 84]. These modules form a nested hierarchy of part-whole relationships [132, 133] that give rise to a bow tie network architecture [119]. Overt behavior reflects the concerted action of large numbers of simple I/O patterns (motor and endocrine primitives) at the bottom of this hierarchy, the activity of which is carefully orchestrated by higher levels of...
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 [60, 63]. In doing so, a hierarchy of generative models is constructed from basic perceptive input (Figure 9).

**Figure 9.** Subfigure (a): The human mind can be modeled as a nested modular hierarchy of generative models that control perception and action [9]. The statistical structure of this hierarchy is assumed to mirror that of the human mental phenotype (i.e. inner experience and overt behavior). Circles indicate generative models and circles within circles submodels. Higher level (domain-general) models are inferred from progressively lower level (domain-specific) models down to the level of perceptive information. Only the (global) flow of prediction errors (red arrows) is shown in this figure (predictions run in the opposite direction, see Figure 6 and 8). Arrows sizes reflect the putative contribution of a particular domain in biasing inference within another domain (see text). Vertical position indicates level of inference. Cognition, emotion, executive functions and motivation occupy a similar level. Within the output hierarchy, higher level generative models impose predictions onto lower level models (blue arrows) to eventually control overt behavior. Domains may affect each other across loops of different path lengths (e.g. from input to output via a hierarchically ordered set of goal states), reflecting different degrees of information processing (evaluation, policy selection). The shortest loops within this hierarchy represent basic stimulus-response patterns (e.g. simple and more complex ‘reflexes’, instinct patterns and instrumental behavior), whereas the longest loops reflect goal-directed behavior that is informed by highly integrated world models involving self-referential, social and normative models (Figure 6).

Each phenotypical domain may have multiple anatomical brain regions as a correlate (see text and references).

**Subfigure (b):** When stressed severely, contextually redundant higher level goal states are shut down to save energy and to enhance the stress response (Figure 8). This corresponds to a collapse of self, social and/or normative models, causing a shift away from goal-directed behavior (longer loops) towards instrumental, habitual, instinctive or reflexive behavior (progressively shorter loops). The loss of higher level constraint triggers a disinhibition of lower levels within the hierarchy, i.e. emotional, motivational and sensorimotor (or sensori-endocrine) mental states. If this change does not force a quick reduction of prediction errors (and corresponding stress responses), this physiological response may revert to a model of psychopathology. Disinhibition may cause hyperconnectivity in control systems (CSD) [107], which strengthens the coupling of an individual with the external environment. This may present as strong interpersonal dependency and/or social conflict.

The statistical structure of this hierarchy is assumed to mirror that of the human mental phenotype (i.e. inner experience and overt behavior; [9], see Figure 9). At the bottom of this hierarchy, three global types of perceptual input can be discerned. Exteroceptive perception involves information coming from the external environment, i.e. the main senses of vision, hearing, touch, smell and taste. Interceptive information involves information feeds coming from the internal environment, e.g. gut and vascular pressure and pain 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 joints. Based on this basic input, the hierarchy of generative models is progressively produced (Figure 9). Recent studies conceptualize human emotion as hierarchical Bayesian infer-
ence on predominantly interoceptive information, placing these generative models somewhere along the middle of this hierarchy [176, 182]. Similarly, cognition may involve hierarchical inference on predominantly exteroceptive information [183]. Executive functions involve part of an output hierarchy that is engaged in high-level (conceptual and premotor) planning, with a predominant connection to proprioceptive input (controlling muscle action) [153]. Motivational functions have been conceptualized as aiding in predicting the precision of motor and endocrine output, perhaps with a predominant influence on endocrine actions [155]. At the top of this hierarchy, humans make highly integrated models of their own inner and outer states along with its history and possible futures (self-image). Such self-models are processed in midline areas of the human brain, which are involved in the highest levels of integrative processing [77, 147, 127]. Additionally, humans make highly integrated models of the inner and outer states of others and their possible histories and futures (theory of mind). Such models involve medioprefrontal areas, temporal poles, the superior temporal sulci, and temporoparietal junction [3, 66, 121]. Like midline self-structures, these areas process information at very high levels of contextual integration. Finally, brain areas have been identified that support a highest level of human functioning that transcends the level of the individual as well as its immediate social environment and connects people across nations and timescales through common laws and values. The existence of such a domain of functioning has been eschewed by scientists for quite some time because of its inherently moral (or even religious) nature. Nevertheless, it is predicted by the principle of hierarchical Bayesian inference (Figure 7). Several empirical studies have found independent components of the human mental phenotype that suggest the presence of normative functions, e.g. Koltko-Rivera [101], Stankov [193], and Collani and Grumm [36]. Individual differences in the degree to which subjects score on these phenotypical dimensions can explain differences in normative or moral behavior. Indeed, empirical studies have demonstrated a taxonomy of goal states that eventually involves the superordinate goals of agency (connecting with the self), communion (to connect with a local social group) and meaning (connecting across spatial, temporal and social barriers) [201]. These superordinate goals are closely related to Maslow’s hierarchy of needs (with multiple layers of self-actualization, social belonging and transcendence; [101]) and have a strong resemblance to self, social and normative or nomothetic goal states as predicted by hierarchical Bayesian inference. Indeed, a large body of literature has identified brain regions involved in normative behavior, which include ventromedial areas for norm processing and right insula, dorsolateral prefrontal, and dorsal cingulate cortices for processing in relation to social norm violation [230]. Such brain areas again involve some of the highest levels of integration across lower level systems. 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 [81]. 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 [47]. 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.

As observed (section 4.3), severe stress seems to modulate the behavior of organisms in a universal way in that it degrades a nested hierarchy of generative models in a top-down fashion. Put bluntly, we propose that severe stress prunes the top of a regulatory pyramid, producing a ‘chicken without a head’ syndrome in any type of organism (a subtle form of ‘decortication’). Although mild forms of stress differentially affect the perception of ourselves and the world around us, severe stress brings us into ‘survival mode’ [29, 117, 213, 125]. Neuroimaging studies show that the human brain falls back from goal-directed to habitual control during stress [175, 174]. This corresponds to decreased activity in higher level systems such as the anterior cingulate, anterio
insular and temporopolar areas [5, 45, 124, 126, 175, 174, 209]. Brain areas that decrease activity during severe stress are midline structures involved in generating self-models (self-image; [70, 87, 97, 192]). Additionally, severe stress decreases activity in brain areas associated with the production of social world models or theory of mind [170, 203]. Finally, severe stress is known to negatively affect moral decision making [106, 129, 194, 224]. This change in behavior is related to altered activity in brain areas involved in transpersonal identification, including law-abiding and moral behavior [106]. Overall, 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, forcing energy to flow across shorter action-perception cycles (Figure 8, 9).

As observed, we propose that a top-down collapse of hierarchical control reduces inhibitory connections, which can lead to a state of critical slowing down (CSD). This phenomenon has been coined as a generic model for mental illness [107]. This idea is supported by recent findings showing that some changes in brain function are common to a diverse range of acute mental disorders (psychopathology, e.g. major unipolar depression, bipolar disorders, psychosis or anxiety disorders). These transdiagnostic changes in functional neuroanatomy include decreased activity in prefrontal and anterior brain regions, which support high-level cognitive control [126]. Acute mental illness can therefore be considered a temporary collapse of higher levels of control that reaches too deeply down the hierarchy of goal states. This causes a dysregulation (disorder) of lower levels within the hierarchy of generative models (Figure 9). These hypotheses can be tested by comparing measures of hierarchical complexity in different brain areas between different stress levels and corresponding behavioral measures, e.g. such changes would result in a compression of the knot of the bow tie structure (Figure 9, see Discussion).

From the above, it follows that individual differences in the degree to which hierarchies of control have grown and matured in the course of life (i.e. personality development) should explain different susceptibilities to mental illness (CSD): 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, neuroimaging studies show that people with (borderline) personality disorders, which are more susceptible to mental decompensation (‘crises’), have low volumes of gray matter in the same areas of cognitive control that are downregulated under stress [28, 126]. This supports the idea that a stress-induced collapse of already underdeveloped regulatory hierarchies triggers CSD in human subjects with personality disorders. Indeed, brain areas that are underdeveloped in people with personality disorders involve the same areas that harbor our world models of self, others and global world views [28]. Thus, it seems that the global faculty of cognitive control that is downregulated in acute mental illness can be subdivided into high-level world models that support agency (self-models), communion (social models) and meaning (normative models), which are each downregulated to different degrees under stress (Figure 9B). These mental faculties may therefore qualify as transdiagnostic factors, which are to some degree involved in all personality disorders (when underdeveloped) and acute mental disorders (when downregulated). Indeed, the ‘alternative model’ for personality disorders in the Diagnostic and Statistical Manual of mental disorders (DSM-5) currently lists self-referential and social functions as two global transdiagnostic factors in personality disorders [229]. These may at some point be supplemented with the third overarching factor (normative/transpersonal functions), a conclusion that is consistent with some existing models of personality development (e.g. [35]). The maturation of these ‘great three’ world models involves a life-long process of associative learning within long stimulus–response loops that have been linked to goal-directed behavior. The development of these mental domains across the lifespan has been referred to as personality development, or more specifically character formation. The various generative models that are subordinate to these three top-level domains (i.e. emotional, motivational, cognitive and executive domains and

Goekoop & De Kleijn | Higher goals from a hierarchical Bayesian control systems perspective 24
subdomains) qualify as lesser-order transdiagnostic factors (Figure 9A). These involve shorter stimulus–response loops that have been associated with habit learning and instrumental learning. Such functions develop at earlier stages of life and their stable expression across years has been referred to as temperament. During episodes of acute stress, this hierarchy collapses in a top-down manner, producing a disinhibition of subordinate levels down to the level of the shortest reflex loops (i.e. instinctive behavior, including a disinhibition of responses such as fight, flight, freezing, reproducing, fainting, freezing, et cetera). Thus, the hierarchy of generative models as shown in Figure 9B reverts to a hierarchical taxonomy of psychopathology (acute mental illness; [103]). The relative contribution of each phenotypical domain to the overall disease presentation can be parsimoniously expressed as a transdiagnostic factor profile.

Apart from explaining individual differences in (susceptibility to) mental illness, the current model may explain individual differences in social interaction. 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 (perception) [62]. Individual differences in hierarchical Bayesian control (e.g. personality) may therefore produce different types of social interaction. Such individual differences determine connectivity at a local level and act as simple rules that knit together complex social networks at a global level. Indeed, social networks may follow similar rules of hierarchical Bayesian inference as shown in Figure 6, i.e. some individuals can be modeled as predictive units and others as units that correct predictions and report on errors. In social networks, however, individuals can be expected to swap roles more flexibly than neurons or molecules, i.e. some individuals may act as predictive agents at some time point and as prediction error agents the next. This produces social clusters in which some opinions and beliefs are held and contrasted with those of other social groups, while trying to get a mutual grip on reality. As observed, low levels of top-down control may induce a state of critical slowing down (CSD, see above) within hierarchical Bayesian control systems. This is true regardless of whether such deficits are temporary (e.g. stress-induced) or chronic (e.g. developmental) in nature. Since CSD represents both within-system and between-system enhanced coupling, two undercontrolled individuals may become strongly coupled. CSD in one person may therefore induce or sustain CSD in another individual, producing highly recursive (clonic) stimulus–response relationships between two individuals. This would be a model of strong mutual dependency and/or intense social conflict, including a mutual loss of law abiding and moral behavior in humans. Like clonic spinal reflexes, vicious cycles in social behavior are a symptom of insufficient higher level control, and typically require a third external party (e.g. mediation, judicial arbitry, or medical intervention) in order to be reduced [50].

5 Discussion
We have shown that living systems can be conceptualized as hierarchical Bayesian control systems with a nested, modular and bottleneck (bow tie) network architecture. Such systems come with specific dynamics that appear universal across species and levels of biological organization. Within input and output areas, nested hierarchies of network clusters allow for perceptual inference and action control, respectively. In order to connect input patterns to adaptive output patterns (policy selection), deep goal hierarchies are produced that incorporate increasing amounts of the inner and outer context of the organism. Such models are continuously updated based on ascending prediction error. Severe stress (prediction error or free energy) causes goal hierarchies to collapse in a top-down manner. We propose that this is due to a shut down of high-level integrative hub structures as a result of insufficient (free) energy dissipation, tilting the balance between functional integration and segregation within nested modular network structures towards functional segregation. A shallow collapse of a goal hierarchy causes output (behavior) to shifts away from long-term and normative towards short-term, self-centered and/or social goals. A deep collapse manifests as
the phenomenon of critical slowing down (CSD) and the occurrence of tipping points. CSD causes a strong coupling between an organism and its environment, equaling a loss of homeostasis. We suggest that the above changes are universal features of living systems and may occur at any spatial scale level of organization. In order to test these predictions, researchers will need to consider whole organisms and social groups and not just parts of them. This has been a major obstacle in the past, but modern data analysis techniques are increasingly becoming closer to studying the full complexity of interactions between genes, proteins, metabolites, neurons, brain areas, animal populations and people (the -omics literature). Such large datasets are now increasingly made available at open databases (e.g. [163, 208]) and can be studied using open analysis platforms (e.g. R or Python) to test predictions. 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 igraph in R [41] or Cytoscape [177]. 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 [88]. 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. [143]). Software has been developed that allows detection of so called ‘rich club’ structures (Figure 4), which are collections of hub nodes that connect significantly more to other hubs than chance levels [150]. Rich clubs involve 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. Rich clubs are an important infrastructure for the nested functional integration as an important ingredient of hierarchical inference (section 3.1). Also, software packages exist that can test network structures for a diverse range of motifs, e.g. [122]. These include bow tie motifs as well as their constituent motifs, such as feedforward loops and feedback loops. At the organism level, we expect biological networks to show a nested bow tie (bottleneck) structure, with cross-connections between different levels of the nested hierarchy that allow for shortcuts, producing processing 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 (Box 1, section 3.2). Such motifs should be counterbalanced by feedback motifs that reflect bottom-up correction of higher level predictions by lower level prediction errors.

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 global flow of energy across nested bow tie structures should involve multiple energy streams converging onto higher level hub structures while ascending in a scale invariant hierarchy, reflecting the functional integration of prediction error signals. Also, it should involve multiple energy streams diverging while descending in such a hierarchy, reflecting top-down predictive control. The throughput areas should involve circular patterns of flow, reflecting the continuous cross-evaluation of incoming information and outgoing responses, i.e. policy selection (i.e. the top-down suppression of prediction errors by high-level predictive models and the adjustment of higher-order predictive models by ascending prediction error). This directionality may be reversed

Goekoop & De Kleijn  |  Higher goals from a hierarchical Bayesian control systems perspective  |  26
in output hierarchies. 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; [222]) 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 driven hubs; [222]) have a large out-degree, referring to the number of outgoing connections that support the process of action control. The throughput parts 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-degree of hubs is 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 prediction error units and top-down predictive 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. [172]). 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 [86] and local [139] directionals of functional connections in the human brain have involved the use of a novel and promising phase transfer entropy measure [111]. 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 directionality into ascending prediction errors and descending 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 [40]. In small world networks, some nodes or clusters only exist by virtue of other nodes or clusters, i.e. they form conditional dependencies in space (a hierarchy of part–whole relationships) [161]. 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. [40]. The authors propose to quantify hierarchy in terms of three key elements, which include treeness (pyramidal shape, or bottleneck structure), feedforwardness (top-down or bottom-up directionals) 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 also controls for the nestedness and directionality of links (causal order, divergence or convergence), but needs to be adapted for weighted networks. Perhaps a more straightforward approach to measuring the
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 [92]. 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 [146]. 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. [105]). Measures of nestedness 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; [167, 205]). 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 [167]. Although this 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 incorporated in 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 [189]. Other measures to quantify information integration and corresponding fitness have been suggested, e.g. Edlund et al. [49].

Together, these measures of (nested) hierarchical order and model complexity can be used to test predictions with respect to the comprehensiveness of hierarchical control in biological networks. We expect the amount of hierarchical 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 smaller (less) and larger (more) amounts of parental investment, solitary (less) and more social (more) behavior, less (less) and more (more) prosocial behavior, 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 extensions than perceptive or output hierarchies. As discussed, the height of a hierarchical tree is related to the ability of an organism to control its internal states and the world around it. Organisms with lesser hierarchies may 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 maturation of control hierarchies can also be linked to individual differences in social interaction. At the macrolevel, social networks may follow the same principles of network structure and function as shown in Figures 6 and 8.

As a general remark, hierarchical Bayesian inference describes a mechanism for inferring ‘signs out of signs’ (e.g. humans as symbolic animals), which amounts to a model of semiotics [53]. 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 [62, 210]. 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 disinformation, which has a strong moral
connotation. 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 [137]. 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. 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, yet are highly successful in predicting events. This relative disconnection may be what is required to finally integrate science and morality safely within a single discipline (the 'moral sciences'; [169]). This being said, it may well be a (categorical) imperative for all people to 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 rules through which people may connect across nations, cultures and timescales. 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 [190]. In this paper, we propose that severe stress alters the network connectivity of biological systems in a universal way, i.e. it should produce a shutdown of hub structures within high level integrative structures of the organism. This reduces the amount of functional integration versus segregation as well as the height of the nested hierarchical tree (see above). Such changes may not cause a significant shift in the small worldness measure, but may affect measures of modularity and, therefore, nested hierarchy. We expect stress to decrease the height of nested hierarchical trees as measured by hierarchical clustering algorithms in areas that harbor high-level goal states. Also, we expect measures of (neural or model) complexity 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). Severe stress should produce shorter, shallower and narrower bow tie motifs, which interferes with the ability of organisms to compress information. This should translate into increasingly smaller loops that run from input via processing to output parts of a 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).

Additionally, we expect stress to alter network connectivity in ways that allow for the categorization of stress-levels into discrete stages, each with distinct and quantifiable features (for a similar attempt, see [166]). ‘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 ([0] Habitual behavior, effortless control). This reflects the successful suppression of low-level ascending and excitatory prediction errors by relatively low-level top-down and inhibitory predictive structures. When stress levels rise to mild or moderate levels, we expect loops of increasing length, which represent the top-down suppression of low-level prediction error by increasingly higher-level generative models to solve more complex challenges ([1] explorative behavior, creative problem solving). This should translate into the activity of increasingly high-level predictive structures with top-down (descending) inhibitory connections to subordinate structures (e.g. high-level FFLs, Box...
In contrast, we expect the activity of higher hierarchical levels (and the corresponding amount of top-down inhibitive connections or FFLs) to decrease again when stress levels become severe. This reflects the dissolution of higher level goal states when the hierarchy is taxed to its limits ([2] emergency or stringent responses, shutdown of integrative hubs, functional segregation). Thus, both low and severe levels of stress should engage shorter loops with corresponding habitual rather than goal-directed forms of behavior. Predictions with respect to changes in the directionality of network connections under different levels of stress can be tested using the quantitative measures described above (e.g. causal modeling and/or phase transfer entropy). The final two stages involve an undercontrolled state of critical slowing down ([3] CSD), which itself predicts a loss of functional or structural integrity ([4] tipping points/decompensation). Both of these stages can be identified objectively [107].

So far, few studies have examined how stress affects measures of hierarchy, (neural or model) complexity and directed connectivity in living organisms. 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 [167, 218]. Another study in post-traumatic stress syndrome reports increased amounts of functional segregation [228]. Yet other studies show that the human brain falls back from goal-directed to habitual control during stress [175, 174]. Such findings are in line with a collapse of high-level integrative control. When stress levels are particularly high, we expect tell-tale signs of undercontrolled control systems in the form of critical slowing down (CSD). As observed, different heights of the hierarchical trees may explain individual differences in the susceptibility to CSD and actual decompensation (tipping points). This can be tested by linking the CSD threshold to measures of hierarchical complexity in different individuals or species. Such studies are readily performed in bacteria and other microbes, where e.g. acidity or salinity levels may be varied to examine bacterial stringent responses and survival [120, 225]. For obvious reasons, however, such studies cannot be easily translated to higher organisms. Actively inducing a state of CSD or decompensation in humans or higher primates would be highly unethical. However, studies of mental illness may provide a natural situation in which to examine individual differences in the susceptibility to CSD (mental illness) in relation to hierarchical complexity. Indeed, a loss of hierarchical organization (high level frontal hubs) has been reported in certain forms of mental illness, e.g. [19]. As discussed in section 4, we expect a stagnation of personality development and the concomitant risk of acute mental illness (psychopathology) to involve a failure of higher-level hierarchical structures to mature properly. A temporary collapse of the neurofunctional hierarchy with respect to a previously attained level of functioning should be a factor common to all episodes of psychopathology. Finally, it would be interesting to examine to what degree the network structure of the human mental phenotype (inner experience and overt behavior) echoes the network structure of hierarchical message passing as shown in Figures 6 and 8. Phenotypical networks indeed show signs of small worldness and nested modular hierarchy (part–whole relationships), which can be explained by physical network architectures capable of hierarchical Bayesian inference. Additionally, one could examine to what degree hierarchical generative models (such as Figure 9) can predict the weight and direction of links in phenotypical networks. 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 affection or social conflict) as well as global phenomena such as innovation and rumor diffusion, voting, migration, strikes, riot behavior, economic slowdown and warfare.

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). Such studies would allow testing of the hypothesis whether cumulative prediction error (excess free energy) eventually shuts down connective hub structures at higher level levels within the nested hierarchy (the knots of the bow ties) when energy demand exceeds energy supply [73]. This should
lead to a top-down dissolution of integrative control, causing the system to fall back on routine rather than goal-directed behavior. The loss of higher level predictive (inhibitive) constraint should produce a state of disinhibition at lower hierarchical levels. Such studies would be a safe way to study the development of CSD and tipping points (decompensation) as a function of hierarchical complexity, providing a general model for individual differences in the susceptibility to disease or death. 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.

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, or for each level of observation within the organism (e.g. genetic, cellular, systems level, or social). Organisms can be described in terms of a limited set of network motifs [4], 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 (Figure 10).

Figure 10. 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 bow tie structure) at different spatial scale levels of observation. Modeling organisms would only require knowledge of the positions and scale levels of a particular type of motif throughout the organism, allowing for significant parsimony of description (i.e. organisms can be ‘compressed’ and ‘decompressed’). The fractal-like structure of biological networks means that this figure may reduce to Figure 1 at the highest spatial scale level of observation (the level of the individual organism) without losing much information. See text for further details.

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 association are significantly improved when allowing for a hierarchical structure of input, throughput and output modules (section 2.2). Since small world networks are non-randomly wired networks that form spontaneously when optimizing the flow of energy through random networks [89], 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. Interestingly, 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 [90]. A further improvement can be made by fitting deep networks with bottleneck (bow tie) structure prior to learning [178]. Such structures 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 [57, 114, 115, 196, 214].

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—most evident in things like variational autoencoders [227]. 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, e.g. they have an explicit hierarchical structure with a certain kind of sparsity. This is particularly evident in the bottleneck or bow tie architectures of variational autoencoders that arguably represent the state-of-the-art in deep learning.

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. 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 seem somewhat far-fetched, 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.

6 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.

References
[1] R. A. Adams, S. Shipp, and K. Friston.
“Predictions not commands: active inference in the motor system”.
In: Brain Structure and Function 218.3 (2013), pp. 611–643. ISSN: 1863-2653.
[2] U. Alon. “Network motifs: theory and experimental approaches”. In: *Nature Reviews Genetics* 8.6 (2007), pp. 450–461. ISSN: 1471-0056 (Print) 1471-0056 (Linking). doi: 10.1038/nrg2102. url: http://www.ncbi.nlm.nih.gov/pubmed/17510665.

[3] D. M. Amodio and C. D. Frith. “Meeting of minds: the medial frontal cortex and social cognition”. In: *Nature Reviews Neuroscience* 7.4 (2006), pp. 268–277. ISSN: 1471-003X (Print) 1471-003X (Linking). doi: 10.1038/nrn1884. url: https://www.ncbi.nlm.nih.gov/pubmed/16552413.

[4] R. P. Araujo and L. A. Liotta. “The topological requirements for robust perfect adaptation in networks of any size”. In: *Nature Communications* 9.1 (2018), p. 1757. ISSN: 2041-1723. doi: 10.1038/s41467-018-04151-6. url: https://doi.org/10.1038/s41467-018-04151-6.

[5] A. F. T. Arnsten. “Stress signalling pathways that impair prefrontal cortex structure and function”. In: *Nature Reviews Neuroscience* 10.6 (2009), pp. 410–422. ISSN: 1471-0048 1471-003X. doi: 10.1038/nrn2648. url: https://www.ncbi.nlm.nih.gov/pubmed/19455173https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2907136/.

[6] W. R. Ashby. *An introduction to cybernetics*. Chapman & Hall, 1961.

[7] W. R. Ashby. “Principles of the Self-Organizing Dynamic System”. In: *Journal of General Psychology* 37.2 (1947), pp. 125–128. ISSN: 0022-1309. doi: 10.1080/00221309.1947.9918144. url: https://doi.org/10.1080/00221309.1947.9918144.

[8] F. A. Azevedo et al. “Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain”. In: *Journal of Comparative Neurology* 513.5 (2009), pp. 532–541. ISSN: 0021-9967. doi: 10.1002/cne.21974.

[9] P. B. Badcock, K. Friston, and M. Ramstead. “The hierarchically mechanistic mind: A free-energy formulation of the human psyche”. In: *Physics of Life Reviews* 31 (2019), pp. 104–121. ISSN: 1571-0645. doi: 10.1016/j.plrev.2018.10.002.

[10] P. Bak, C. Tang, and K. Wiesenfeld. “Self-organized criticality”. In: *Physical Review A* 38.1 (1988), pp. 364–374. doi: 10.1103/PhysRevA.38.364.

[11] C. L. Baker et al. “Rational quantitative attribution of beliefs, desires and percepts in human mentalizing”. In: *Nature Human Behaviour* 1 (2017), p. 64. doi: 10.1038/s41562-017-0064.

[12] A. L. Barabasi. “Network science”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (2013), p. 20120375. ISSN: 1364-503X (Print) 1364-503X (Linking). doi: 10.1098/rsta.2012.0375. url: http://www.ncbi.nlm.nih.gov/pubmed/23419844.

[13] A. L. Barabasi. “Scale-free networks: a decade and beyond”. In: *Science* 325.5939 (2009), pp. 412–413. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1173299. url: http://www.ncbi.nlm.nih.gov/pubmed/19628854.

[14] A. L. Barabasi and E. Bonabeau. “Scale-free networks”. In: *Scientific American* 288.5 (2003), pp. 60–69. url: PM:12701331.

[15] A. L. Barabasi and Z. N. Oltvai. “Network biology: understanding the cell’s functional organization”. In: *Nature Reviews Genetics* 5 (2004), pp. 101–113. doi: 10.1038/nrg1272.

[16] B. Barzel and A. L. Barabasi. “Network link prediction by global silencing of indirect correlations”. In: *Nature Biotechnology* 31.8 (2013), pp. 720–725. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking). doi: 10.1038/nbt.2601. url: http://www.ncbi.nlm.nih.gov/pubmed/23851447.

[17] B. Barzel and A. L. Barabasi. “Universality in network dynamics”. In: *Nature Physics* 9 (2013), pp. 673–681. ISSN: 1745-2473 (Print) 1745-2473 (Linking). doi: 10.1038/nphys2741. url: http://www.ncbi.nlm.nih.gov/pubmed/24319492.

[18] D. S. Bassett and E. D. Bullmore. “Small-world brain networks”. In: *Neuroscientist* 12.6 (2006), pp. 512–523. ISSN: 1073-8584. doi: 10.1177/1073858406293182.
[19] D. S. Bassett et al. “Hierarchical organization of human cortical networks in health and schizophrenia”. In: Journal of Neuroscience 28.37 (2008), pp. 9239–9248. doi: 10.1523/JNEUROSCI.1929-08.2008.

[20] M. Bekoff and J. Pierce. Wild justice: The moral lives of animals. University of Chicago Press, 2009. ISBN: 0226041662.

[21] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. “Variational inference: A review for statisticians”. In: Journal of the American Statistical Association 112.518 (2017), pp. 859–877. ISSN: 0162-1459. doi: 10.1080/01621459.2017.1285773.

[22] M. M. Botvinick. “Hierarchical models of behavior and prefrontal function”. In: Trends in Cognitive Sciences 12.5 (2008), pp. 201–208. ISSN: 1364-6613 (Print) 1364-6613 (Linking). doi: 10.1016/j.tics.2008.02.009. URL: https://www.ncbi.nlm.nih.gov/pubmed/18420448.

[23] M. M. Botvinick. “Multilevel structure in behaviour and in the brain: a model of Fuster’s hierarchy”. In: Philosophical Transactions of the Royal Society B: Biological Sciences 362.1485 (2007), pp. 1615–1626. ISSN: 0962-8436 (Print) 0962-8436 (Linking). doi: 10.1098/rstb.2007.2056. URL: https://www.ncbi.nlm.nih.gov/pubmed/17428777.

[24] M. M. Botvinick and A. Weinstein. “Model-based hierarchical reinforcement learning and human action control”. In: Philosophical Transactions of the Royal Society B: Biological Sciences 369.1655 (2014). ISSN: 1471-2970 (Electronic) 0962-8436 (Linking). doi: 10.1098/rstb.2013.0480. URL: https://www.ncbi.nlm.nih.gov/pubmed/25267822.

[25] V. Braitenberg. Vehicles: Experiments in synthetic psychology. Cambridge MA: MIT Press, 1984.

[26] B. Brembs. “Operant conditioning in invertebrates”. In: Current Opinion in Neurobiology 13.6 (2003), pp. 710–717. ISSN: 0959-4388. doi: 10.1016/j.conb.2003.10.002.

[27] R. Brooks. “A robust layered control system for a mobile robot”. In: IEEE Journal of Robotics and Automation 2.1 (1986), pp. 14–23. doi: 10.1109/JRA.1986.1087032.

[28] R. Brunner et al. “Reduced prefrontal and orbitofrontal gray matter in female adolescents with borderline personality disorder: is it disorder specific?” In: NeuroImage 49.1 (2010), pp. 114–120. ISSN: 1053-8119. doi: 10.1016/j.neuroimage.2009.07.070.

[29] T. W. Buchanan and S. D. Preston. “Stress leads to prosocial action in immediate need situations”. In: Frontiers in Behavioral Neuroscience 8 (2014), p. 5. ISSN: 1662-5153. doi: 10.3389/fnbeh.2014.00005.

[30] E. Bullmore and O. Sporns. “Complex brain networks: graph theoretical analysis of structural and functional systems”. In: Nature Reviews Neuroscience 10.3 (2009), pp. 186–198. doi: 10.1038/nrn2575. URL: PM:19190637.

[31] P. Calvo and F. Baluška. “Conditions for minimal intelligence across eukaryota: a cognitive science perspective”. In: Frontiers in Psychology 6 (2015), p. 1329. ISSN: 1664-1078. doi: 10.3389/fpsyg.2015.01329.

[32] P. Calvo and K. Friston. “Predicting green: really radical (plant) predictive processing”. In: Journal of The Royal Society Interface 14.131 (2017), p. 20170096. ISSN: 1742-5689. doi: 10.1098/rsif.2017.0096.

[33] W. B. Cannon. “Organization for physiological homeostasis”. In: Physiological Reviews 9.3 (1929), pp. 399–431. ISSN: 0031-9333. doi: 10.1152/physrev.1929.9.3.399.

[34] W. B. Cannon. The wisdom of the body. Norton, 1932.

[35] C. R. Cloninger. “The psychobiological theory of temperament and character: comment on Farmer and Goldberg (2008)”. In: Psychological Assessment 20.3 (2008), 292–9; discussion 300–4. ISSN: 1040-3590 (Print) 1040-3590 (Linking). doi: 10.1037/a0012933. URL: http://www.ncbi.nlm.nih.gov/pubmed/18778165.

[36] G. von Collani and M. Grumm. “On the dimensional structure of personality, ideological beliefs, social attitudes, and personal values”. In: Journal of Individual Differences 30.2 (2009), pp. 107–119. doi: 10.1027/1614-0001.30.2.107.
[37] M. L. Commons and A. Pekker. “Presenting the Formal Theory of Hierarchical Complexity”. In: World Futures 64.5-7 (2008), pp. 375–382. issn: 0260-4027. doi: 10.1080/02604020802301204. url: https://doi.org/10.1080/02604020802301204.

[38] A. Constant et al. “A variational approach to niche construction”. In: Journal of The Royal Society Interface 15.141 (2018), p. 20170685. issn: 1742-5689. doi: 10.1098/rsif.2017.0685.

[39] A. Constant et al. “Regimes of expectations: An active inference model of social conformity and decision making”. In: Frontiers in Psychology 10 (2019), p. 679. issn: 1664-1078. doi: 10.3389/fpsyg.2019.00679.

[40] B. Corominas-Murtra et al. “On the origins of hierarchy in complex networks”. In: Proceedings of the National Academy of Sciences of the United States of America 110.33 (2013), pp. 13316–13321. issn: 0027-8424. doi: 10.1073/pnas.1300832110.

[41] G. Csardi and T. Nepusz. “The igraph software package for complex network research”. In: InterJournal 1695.5 (2005), pp. 1–9.

[42] M. Csete and J. Doyle. “Bow ties, metabolism and disease”. In: Trends in Biotechnology 22.9 (2004), pp. 446–450. issn: 0167-7799. doi: 10.1016/j.tibtech.2004.07.007.

[43] M. Del Giudice, S. W. Gangestad, and H. S. Kaplan. “Life history theory and evolutionary psychology”. In: The handbook of evolutionary psychology. Vol. 1: Foundations (2015), pp. 88–114. doi: 10.1002/9781119125563.evpsy1ch102.

[44] M. Del Giudice et al. “What is stress? A systems perspective”. In: Integrative and Comparative Biology 58.6 (2018), pp. 1019–1032. issn: 1540-7063. doi: 10.1093/icb/icy114.

[45] E. Dias-Ferreira et al. “Chronic Stress Causes Frontostriatal Reorganization and Affects Decision-Making”. In: Science 325.5940 (2009), pp. 621–625. doi: 10.1126/science.1171203. url: http://science.sciencemag.org/content/325/5940/621.abstract.

[46] B. B. Doll, D. A. Simon, and N. D. Daw. “The ubiquity of model-based reinforcement learning”. In: Current Opinion in Neurobiology 22.6 (2012), pp. 1075–1081. issn: 0959-4388. doi: 10.1016/j.conb.2012.08.003.

[47] R. I. Dunbar and S. Shultz. “Evolution in the social brain”. In: Science 317.5843 (2007), pp. 1344–1347. issn: 0036-8075. doi: 10.1126/science.1145463.

[48] T. Durfee et al. “Transcription profiling of the stringent response in Escherichia coli”. In: Journal of Bacteriology 190.3 (2008), pp. 1084–1096. issn: 0021-9193. doi: 10.1128/JB.01092-07.

[49] J. A. Edlund et al. “Integrated information increases with fitness in the evolution of animals”. In: PLoS Computational Biology 7.10 (2011), e1002236. issn: 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1002236. url: https://www.ncbi.nlm.nih.gov/pubmed/22028639.

[50] E. Fehr and U. Fischbacher. “Third-party punishment and social norms”. In: Evolution and Human Behavior 25.2 (2004), pp. 63–87. issn: 1090-5138. doi: 10.1016/S1090-5138(04)00005-4.

[51] E. Fehr and I. Schurtenberger. “Normative foundations of human cooperation”. In: Nature Human Behaviour 2 (2018), pp. 458–468. issn: 2397-3374. doi: 10.1038/s41562-018-0385-5.

[52] A. G. Feldman and M. F. Levin. “The equilibrium-point hypothesis—past, present and future”. In: Progress in motor control. Springer, 2009, pp. 699–726. doi: 10.1007/978-0-387-77064-2_38.

[53] M. Fortier and D. A. Friedman. “Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston”. In: ALIUS Bulletin 2 (2018), pp. 17–43. url: https://www.aliusresearch.org/uploads/9/1/6/0/91600416/friston___of_woodlice_and_men.pdf.

[54] W. J. Freeman. “A field-theoretic approach to understanding scale-free neural dynamics”. In: Biological Cybernetics 92.6 (2005), pp. 350–359. issn: 0340-1200. doi: 10.1007/s00422-005-0563-1.
[55] W. J. Freeman. *How brains make up their minds*. Columbia University Press, 2001.

[56] T. Friedlander et al. “Evolution of bow-tie architectures in biology”. In: *PLoS Computational Biology* 11.3 (2015), e1004055. issn: 1553-7358. doi: 10.1371/journal.pcbi.1004055.

[57] K. Friston. “A free energy principle for a particular physics”. In: *arXiv:1906.10184 [q-bio.NC]* (2019). url: https://arxiv.org/abs/1906.10184.

[58] K. Friston. “A free energy principle for biological systems”. In: *Entropy* 14.11 (2012), pp. 2100–2121. doi: 10.3390/e14112100.

[59] K. Friston. “Does predictive coding have a future?” In: *Nature Neuroscience* 21 (2018), pp. 1019–1021. issn: 1546-1726. doi: 10.1038/s41593-018-0200-7.

[60] K. Friston. “The free-energy principle: a unified brain theory?” In: *Nature Reviews Neuroscience* 11 (2010), pp. 127–138. doi: 10.1038/nrn2787.

[61] K. Friston. “Waves of prediction”. In: *PLoS Biology* 17 (2019), e3000426. doi: 10.1371/journal.pbio.3000426.

[62] K. Friston and C. D. Frith. “Active inference, communication and hermeneutics”. In: *Cortex* 68 (2015), pp. 129–143. issn: 0010-9452. doi: 10.1016/j.cortex.2015.03.025.

[63] K. Friston, J. Kilner, and L. Harrison. “A free energy principle for the brain”. In: *Journal of Physiology-Paris* 101.1–3 (2006), pp. 70–87. issn: 0928-4257. doi: 10.1016/j.physparis.2006.10.001.

[64] K. Friston, T. Parr, and B. de Vries. “The graphical brain: belief propagation and active inference”. In: *Network Neuroscience* 1.4 (2017), pp. 381–414. issn: 2472-1751. doi: 10.1162/NETN_a_00018.

[65] K. Friston et al. “The anatomy of choice: active inference and agency”. In: *Frontiers in Human Neuroscience* 7.598 (2013), issn: 1662-5161. doi: 10.3389/fnhum.2013.00598. url: https://www.frontiersin.org/article/10.3389/fnhum.2013.00598.

[66] H. L. Gallagher and C. D. Frith. “Functional imaging of ‘theory of mind’”. In: *Trends in Cognitive Sciences* 7.2 (2003), pp. 77–83. issn: 1364-6613. doi: 10.1016/S1364-6613(02)00025-6.

[67] L. K. Gallos, C. Song, and H. A. Makse. “A review of fractality and self-similarity in complex networks”. In: *Physica A: Statistical Mechanics and its Applications* 386.2 (2007), pp. 686–691. issn: 0378-4371. doi: 10.1016/j.physa.2007.07.069.

[68] M. Girvan and M. E. Newman. “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12 (2002), pp. 7821–7826. issn: 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.122653799. url: http://www.ncbi.nlm.nih.gov/pubmed/12060727.

[69] A. Goelzer and V. Fromion. “Resource allocation in living organisms”. In: *Biochemical Society Transactions* 45.4 (2017), pp. 945–952. issn: 0300-5127. doi: 10.1042/bst20160436. url: https://doi.org/10.1042/BST20160436.

[70] L. Goette et al. “Stress pulls us apart: Anxiety leads to differences in competitive confidence under stress”. In: *Psychoneuroendocrinology* 54 (2015), pp. 115–123. issn: 0306-4530. doi: https://doi.org/10.1016/j.psyneuen.2015.01.019. url: http://www.sciencedirect.com/science/article/pii/S0306453015000335.

[71] E. E. Gorenstein and J. P. Newman. “Disinhibitory psychopathology: A new perspective and a model for research”. In: *Psychological Review* 87.3 (1980), pp. 301–315. issn: 1939-1471. doi: 10.1037/0033-295X.87.3.301.

[72] M. Gosak et al. “Network science of biological systems at different scales: a review”. In: *Physics of Life Reviews* 24 (2018), pp. 118–135. doi: 10.1016/j.plrev.2017.11.003.
[73] M. Gosak et al. “The relationship between node degree and dissipation rate in networks of diffusively coupled oscillators and its significance for pancreatic beta cells”. In: Chaos 25.7 (2015), p. 073115. doi: 10.1063/1.4926673.

[74] T. L. Griffiths et al. “Probabilistic Models of Cognition: Exploring the laws of thought”. In: Trends in Cognitive Sciences (2010), pp. 357–364. doi: 10.1016/j.tics.2010.05.004.

[75] T. Gross and B. Blasius. “Adaptive coevolutionary networks: a review”. In: Journal of the Royal Society Interface 5.20 (2007), pp. 259–271. ISSN: 1742-5689. doi: 10.1098/rsif.2007.1229.

[76] H. Hafez-Kolahi and S. Kasaei. “Information Bottleneck and its Applications in Deep Learning”. In: arXiv:1904.03743 [cs.LG] (2019).

[77] P. Haggard. “Sense of agency in the human brain”. In: Nature Reviews Neuroscience 18 (2017), pp. 196–207. ISSN: 1471-0048. doi: 10.1038/nrn.2017.14.

[78] G. Hausfater and S. B. Hrdy. Infanticide: comparative and evolutionary perspectives. Routledge, 2017. ISBN: 1351512625.

[79] J. Hegdé and D. J. Felleman. “Reappraising the functional implications of the primate visual anatomical hierarchy”. In: Neuroscientist 13.5 (2007), pp. 416–421. ISSN: 1073-8584. doi: 10.1177/1073858407305201.

[80] E. J. Hermans et al. “Dynamic adaptation of large-scale brain networks in response to acute stressors”. In: Trends in Neurosciences 37.6 (2014), pp. 304–314. ISSN: 1878-108X (Electronic) 0166-2236 (Linking). doi: 10.1016/j.tins.2014.03.006. URL: https://www.ncbi.nlm.nih.gov/pubmed/24766931.

[81] E. Herrmann et al. “Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis”. In: Science 317.5843 (2007), pp. 1360–1366. doi: 10.1126/science.1146282.

[82] C. Hesp et al. “A multi-scale view of the emergent complexity of life: A free-energy proposal”. In: Evolution, Development and Complexity. Springer, 2019, pp. 195–227. doi: 10.1007/978-3-030-00075-2_7.

[83] M. P. van den Heuvel et al. “High-cost, high-capacity backbone for global brain communication”. In: Proceedings of the National Academy of Sciences of the United States of America 109 (28 2012), pp. 11372–11377. doi: 10.1073/pnas.120393109.

[84] M. P. van den Heuvel et al. “Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain”. In: NeuroImage 43.3 (2008), pp. 528–539. doi: 10.1016/j.neuroimage.2008.08.010.

[85] C. C. Hilgetag and A. Goulas. “Is the brain really a small-world network?” In: Brain Structure and Function 221 (2016), pp. 2361–2366. ISSN: 1863-2661 (Electronic) 1863-2653 (Linking). doi: 10.1007/s00429-015-1035-6. URL: https://www.ncbi.nlm.nih.gov/pubmed/25894630.

[86] A. Hillebrand et al. “Direction of information flow in large-scale resting-state networks is frequency-dependent”. In: Proceedings of the National Academy of Sciences of the United States of America 113.14 (2016), pp. 3867–3872. DOI: 10.1073/pnas.1515657113. URL: http://www.pnas.org/content/113/14/3867.abstract.

[87] J. M. Hooley et al. “Activation in dorsolateral prefrontal cortex in response to maternal criticism and praise in recovered depressed and healthy control participants”. In: Biological Psychiatry 57.7 (2005), pp. 809–812. doi: 10.1016/j.biopsych.2005.01.012.

[88] M. D. Humphries and K. Gurney. “Network ‘Small-World-Ness’: a quantitative method for determining canonical network equivalence”. In: PLoS ONE 3.4 (2008), e0002051. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0002051. URL: http://www.ncbi.nlm.nih.gov/pubmed/18446219.
[89] N. Jarman et al. “Self-organisation of small-world networks by adaptive rewiring in response to graph diffusion”. In: *Scientific Reports* 7.1 (2017), p. 13158. ISSN: 2045-2322. doi: 10.1038/s41598-017-12589-9. URL: https://doi.org/10.1038/s41598-017-12589-9.

[90] M. Javaheripi, B. D. Rouhani, and F. Koushanfar. “SWNet: Small-World Neural Networks and Rapid Convergence”. In: *arXiv:1904.04862 [cs.LG]* (2019). URL: https://arxiv.org/abs/1904.04862.

[91] M. I. Jordan and T. M. Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), p. 255. doi: 10.1126/science.aa8415. URL: http://science.sciencemag.org/content/349/6245/255.abstract.

[92] M. Kaiser and C. Hilgetag. “Optimal hierarchical modular topologies for producing limited sustained activation of neural networks”. In: *Frontiers in Neuroinformatics* 4 (2010), p. 8. ISSN: 1662-5196. URL: https://www.frontiersin.org/article/10.3389/fninf.2010.00008.

[93] R. Kanai et al. “Cerebral hierarchies: predictive processing, precision and the pulvinar”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1668 (2015), p. 20140169. doi: 10.1098/rstb.2014.0169.

[94] Y. Karklin and M. S. Lewicki. “Emergence of complex cell properties by learning to generalize in natural scenes”. In: *Nature* 457.7225 (2009), pp. 83–86. doi: 10.1038/nature07481.

[95] S. A. Kauffman. *At home in the universe: The search for the laws of self-organization and complexity*. Oxford University Press, 1996. ISBN: 019976185X.

[96] S. A. Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, 1993. ISBN: 0195079515.

[97] M.-L. Kesting et al. “The impact of social stress on self-esteem and paranoid ideation”. In: *Journal of Behavior Therapy and Experimental Psychiatry* 44.1 (2013), pp. 122–128. doi: 10.1016/j.jbtep.2012.07.010.

[98] H. Kitano. “Biological Complexity and the Need for Computational Approaches”. In: *Philosophy of Systems Biology*. Springer, 2017, pp. 169–180. doi: 10.1007/978-3-319-47000-9_16.

[99] H. Kitano. “Biological robustness”. In: *Nature Reviews Genetics* 5.11 (2004), pp. 826–837. doi: 10.1038/nrg1471.

[100] R. de Kleijn, G. Kachergis, and B. Hommel. “Everyday robotic action: lessons from human action control”. In: *Frontiers in Neurorobotics* 8 (2014), p. 13. ISSN: 1662-5218. doi: 10.3389/fnbot.2014.00013.

[101] M. E. Koltko-Rivera. “Rediscovering the later version of Maslow’s hierarchy of needs: Self-transcendence and opportunities for theory, research, and unification”. In: *Review of General Psychology* 10.4 (2006), pp. 302–317. doi: 10.1037/1089-2680.10.4.302.

[102] J. M. Koolhaas et al. “Stress revisited: a critical evaluation of the stress concept”. In: *Neuroscience & Biobehavioral Reviews* 35.5 (2011), pp. 1291–1301. doi: 10.1016/j.neubiorev.2011.02.003.

[103] R. Kotov et al. “The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies”. In: *Journal of Abnormal Psychology* 126.4 (2017), p. 454. doi: 10.1037/abn0000258.

[104] N. Kriegeskorte. “Deep neural networks: a new framework for modeling biological vision and brain information processing”. In: *Annual Review of Vision Science* 1 (2015), pp. 417–446. doi: 10.1146/annurev-vision-082114-035447.

[105] A. Lancichinetti and S. Fortunato. “Community detection algorithms: a comparative analysis”. In: *Physical Review E* 80.5 (2009), p. 056117. doi: 10.1103/PhysRevE.80.056117.

[106] E.-J. Lee and J. H. Yun. “Moral incompetency under time constraint”. In: *Journal of Business Research* 99 (2019), pp. 438–445. doi: 10.1016/j.jbusres.2017.10.043.
[107] I. A. van de Leemput et al.  “Critical slowing down as early warning for the onset and termination of depression”. In: Proceedings of the National Academy of Sciences of the United States of America 111.1 (2014), pp. 87–92. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1312114110. url: http://www.ncbi.nlm.nih.gov/pubmed/24324144.

[108] J. Li et al. “The architecture of the gene regulatory networks of different tissues”. In: Bioinformatics 28.18 (2012), pp. i509–i514. doi: 10.1093/bioinformatics/bts387.

[109] J. Limanowski and F. Blankenburg, “Minimal self-models and the free energy principle”. In: Frontiers in Human Neuroscience 7 (2013), p. 547. doi: 10.3389/fnhum.2013.00547.

[110] G. Litijens et al. “A survey on deep learning in medical image analysis”. In: Medical Image Analysis 42 (2017), pp. 60–88. doi: 10.1016/j.media.2017.07.005.

[111] M. Lobier et al. “Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions”. In: NeuroImage 85 (2014), pp. 853–872. doi: 10.1016/j.neuroimage.2013.08.056.

[112] P. Lyon. “The cognitive cell: bacterial behavior reconsidered”. In: Frontiers in Microbiology 6 (2015), p. 264. issn: 1664-302X. doi: 10.3389/fmicb.2015.00264. url: https://www.ncbi.nlm.nih.gov/pubmed/25928619https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4396460/.

[113] L. López-Maury, S. Marguerat, and J. Bähler. “Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation”. In: Nature Reviews Genetics 9.8 (2008), pp. 583–593. doi: 10.1038/nrg2398.

[114] D. J. C. MacKay. “Free energy minimisation algorithm for decoding and cryptanalysis”. In: Electronics Letters 31.6 (1995), pp. 446–447. doi: 10.1049/el:19950331.

[115] D. J. C. MacKay. Information theory, inference, and learning algorithms. Cambridge University Press, 2003. ISBN: 0521642981.

[116] D. Maisto, F. Donnarumma, and G. Pezzulo. “Divide et impera: subgoaling reduces the complexity of probabilistic inference and problem solving”. In: Journal of the Royal Society Interface 12.104 (2015), p. 20141335. doi: 10.1098/rsif.2014.1335.

[117] Y. Mao et al. “Optimal experience and optimal identity: A multinational study of the associations between flow and social identity”. In: Frontiers in Psychology 7 (2016), p. 67. doi: 10.3389/fpsyg.2016.00067.

[118] G. Marchetti et al. “A feedforward–feedback glucose control strategy for type 1 diabetes mellitus”. In: Journal of Process Control 18.2 (2008), pp. 149–162. doi: 10.1016/j.jprocont.2007.07.008.

[119] N. T. Markov et al. “Cortical high-density counterstream architectures”. In: Science 342.6158 (2013), p. 1238406. doi: 10.1126/science.1238406.

[120] J. Marles-Wright et al. “Molecular architecture of the “stressosome,” a signal integration and transduction hub”. In: Science 322.5898 (2008), pp. 92–96. doi: 10.1126/science.1159572.

[121] R. B. Mars et al. “Connectivity profiles reveal the relationship between brain areas for social cognition in human and monkey temporoparietal cortex”. In: Proceedings of the National Academy of Sciences of the United States of America 110.26 (2013), pp. 10806–10811. doi: 10.1073/pnas.1302956110.

[122] A. Masoudi-Nejad, F. Schreiber, and Z. R. M. Kashani. “Building blocks of biological networks: a review on major network motif discovery algorithms”. In: IET Systems Biology 6.5 (2012), pp. 164–174. doi: 10.1049/iet-syb.2011.0011.

[123] J. L. McClelland et al. “Letting structure emerge: connectionist and dynamical systems approaches to cognition”. In: Trends in Cognitive Sciences 14.8 (2010), pp. 348–356. doi: 10.1016/j.tics.2010.06.002.
[124] B. S. McEwen et al. “Mechanisms of stress in the brain”. In: Nature Neuroscience 18.10 (2015), pp. 1353–63. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking). doi: 10.1038/nn.4086. url: https://www.ncbi.nlm.nih.gov/pubmed/26404710.

[125] B. S. McEwen and J. C. Wingfield. “The concept of allostasis in biology and biomedicine”. In: Hormones and Behavior 43.1 (2003), pp. 2–15. ISSN: 0018-506X. doi: https://doi.org/10.1016/S0018-506X(02)00024-7. url: http://www.sciencedirect.com/science/article/pii/S0018506X02000247.

[126] L. M. McTeague, M. S. Goodkind, and A. Etkin. “Transdiagnostic impairment of cognitive control in mental illness”. In: Journal of Psychiatric Research 83 (2016), pp. 37–46. doi: 10.1016/j.jpsychires.2016.08.001.

[127] L. van der Meer et al. “Self-reflection and the brain: a theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia”. In: Neuroscience & Biobehavioral Reviews 34.6 (2010), pp. 935–946. doi: 10.1016/j.neubiorev.2009.12.004.

[128] A. J. Meeske et al. “High-Throughput Genetic Screens Identify a Large and Diverse Collection of New Sporulation Genes in Bacillus subtilis”. In: PLOS Biology 14.1 (2016), e1002341. doi: 10.1371/journal.pbio.1002341. url: https://doi.org/10.1371/journal.pbio.1002341.

[129] M. F. Mendez. “The neurobiology of moral behavior: review and neuropsychiatric implications”. In: CNS Spectrums 14.11 (2009), pp. 608–620. doi: 10.1017/S1092852900023853.

[130] M. Mesulam. “From sensation to cognition”. In: Brain 121.6 (1998), pp. 1013–1052. doi: 10.1093/brain/121.6.1013.

[131] M. Mesulam. “Representation, inference, and transcendent encoding in neurocognitive networks of the human brain”. In: Annals of Neurology 64.4 (2008), pp. 367–378. doi: 10.1002/ana.21534.

[132] D. Meunier, R. Lambiotte, and E. T. Bullmore. “Modular and hierarchically modular organization of brain networks”. In: Frontiers in Neuroscience 4 (2010), p. 200. doi: 10.3389/fnins.2010.00200.

[133] D. Meunier et al. “Hierarchical modularity in human brain functional networks”. In: Frontiers in Neuroinformatics 3 (2009), p. 37. doi: 10.3389/neuro.11.037.2009.

[134] S. Milgram. “The small world problem”. In: Psychology Today 1.1 (1967), pp. 61–67.

[135] A. Mitchell et al. “Adaptive prediction of environmental changes by microorganisms”. In: Nature 460 (2009), pp. 220–224. doi: 10.1038/nature08112.

[136] V. Mnih et al. “Human-level control through deep reinforcement learning”. In: Nature 518 (2015), pp. 529–533. doi: 10.1038/nature14236.

[137] G. E. Moore and T. Baldwin. Principia ethica. Cambridge University Press, 1993. ISBN: 0521448484.

[138] M. Moutoussis et al. “Bayesian inferences about the self (and others): A review”. In: Consciousness and Cognition 25 (2014), pp. 67–76. doi: 10.1016/j.concog.2014.01.009.

[139] M. Märtens et al. “Brain network clustering with information flow motifs”. In: Applied Network Science 2.1 (2017), p. 25. ISSN: 2364-8228. doi: 10.1007/s41109-017-0046-z. url: https://doi.org/10.1007/s41109-017-0046-z.

[140] E. de Nadal, G. Ammerer, and F. Posas. “Controlling gene expression in response to stress”. In: Nature Reviews Genetics 12.12 (2011), pp. 833–845. ISSN: 1471-0064. doi: 10.1038/nrg3055. url: https://doi.org/10.1038/nrg3055.

[141] A. Nagabandi et al. “Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning”. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). 2018, pp. 7559–7566. doi: 10.1109/ICRA.2018.8463189.

[142] S. D. Nagar et al. “A Network Biology Approach to Decipher Stress Response in Bacteria Using Escherichia coli As a Model”. In: OMICS 20.5 (2016), pp. 310–324. doi: 10.1089/omi.2016.0028.
M. E. Newman. “Analysis of weighted networks”. In: *Physical Review E* 70.5 Pt 2 (2004), p. 056131. doi: 10.1103/PhysRevE.70.056131.

M. E. Newman. “Finding community structure in networks using the eigenvectors of matrices”. In: *Physical Review E* 74.3 Pt 2 (2006), p. 036104. doi: 10.1103/PhysRevE.74.036104.

M. E. Newman, A. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

M. E. Newman and M. Girvan. “Finding and evaluating community structure in networks”. In: *Physical Review E* 69.2 Pt 2 (2004), p. 026113. doi: 10.1103/PhysRevE.69.026113.

G. Northoff et al. “Self-referential processing in our brain—A meta-analysis of imaging studies on the self”. In: *NeuroImage* 31.1 (2006), pp. 440–457. issn: 1053-8119. doi: 10.1016/j.neuroimage.2005.12.002.

Z. N. Oltvai and A. L. Barabasi. “Systems biology. Life’s complexity pyramid”. In: *Science* 298.5594 (2002), pp. 763–764. doi: 10.1126/science.1078563.

S. Ondobaka, J. Kilner, and K. Friston. “The role of interoceptive inference in theory of mind”. In: *Brain and Cognition* 112 (2017), pp. 64–68. doi: 10.1016/j.bandc.2015.08.002.

T. Opsahl et al. “Prominence and control: the weighted rich-club effect”. In: *Physical Review Letters* 101.16 (2008), p. 168702. doi: 10.1103/PhysRevLett.101.168702.

R. Pascanu et al. “Learning model-based planning from scratch”. In: *arXiv*:1707.06170 [cs.AI] (2017). url: https://arxiv.org/abs/1707.06170.

A. Peters, B. S. McEwen, and K. Friston. “Uncertainty and stress: Why it causes diseases and how it is mastered by the brain”. In: *Progress in Neurobiology* 156 (2017), pp. 164–188. issn: 0301-0082. doi: 10.1016/j.pneurobio.2017.05.004.

G. Pezzulo. “An Active Inference view of cognitive control”. In: *Frontiers in Psychology* 3 (2012), p. 478. doi: 10.3389/fpsyg.2012.00478.

G. Pezzulo, F. Rigoli, and K. Friston. “Active Inference, homeostatic regulation and adaptive behavioural control”. In: *Progress in Neurobiology* 134 (2015), pp. 17–35. doi: 10.1016/j.pneurobio.2015.09.001.

G. Pezzulo, F. Rigoli, and K. Friston. “Hierarchical active inference: a theory of motivated control”. In: *Trends in Cognitive Sciences* 22.4 (2018), pp. 294–306. doi: 10.1016/j.tics.2018.01.009.

W. T. Powers. *Behavior: The control of perception*. Aldine, 1973. isbn: 0202251136.

W. T. Powers. “Feedback: Beyond Behaviorism”. In: *Science* 179.4071 (1973), pp. 351–356. doi: 10.1126/science.179.4071.351.

W. T. Powers, R. Clark, and R. McFarland. “A general feedback theory of human behavior: Part II”. In: *Perceptual and Motor Skills* 11.3 (1960), pp. 309–323. doi: 10.2466/pms.1960.11.3.309.

S. Racanière et al. “Imagination-augmented agents for deep reinforcement learning”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 5690–5701.

M. J. D. Ramstead, P. B. Badcock, and K. Friston. “Answering Schrödinger’s question: A free-energy formulation”. In: *Physics of Life Reviews* 24 (2018), pp. 1–16. doi: 10.1016/j.plrev.2017.09.001.

E. Ravasz and A. L. Barabasi. “Hierarchical organization in complex networks”. In: *Physical Review E* 67.2 Pt 2 (2003), p. 026122. doi: 10.1103/PhysRevE.67.026122.

J. J. Ribas-Fernandes et al. “A neural signature of hierarchical reinforcement learning”. In: *Neuron* 71.2 (2011), pp. 370–9. issn: 1097-4199 (Electronic) 0896-6273 (Linking). doi: 10.1016/j.neuron.2011.05.042. url: https://www.ncbi.nlm.nih.gov/pubmed/21791234.

D. J. Rigden and X. M. Fernández. “The 2018 Nucleic Acids Research database issue and the online molecular biology database collection”. In: *Nucleic Acids Research* 46.D1 (2017), pp. D1–D7. issn: 0305-1048. doi: 10.1093/nar/gkx1235. url: https://doi.org/10.1093/nar/gkx1235.
[164] T. Rohe, A.-C. Ehlis, and U. Noppeney. “The neural dynamics of hierarchical Bayesian causal inference in multisensory perception”. In: Nature Communications 10 (2019), pp. 1–17. doi: 10.1038/s41467-019-09664-2.

[165] T. Rohe and U. Noppeney. “Cortical hierarchies perform Bayesian causal inference in multisensory perception”. In: PLoS Biology 13.2 (2015). doi: 10.1371/journal.pbio.1002073.

[166] L. M. Romero, M. J. Dickens, and N. E. Cyr. “The Reactive Scope Model—a new model integrating homeostasis, allostatics, and stress”. In: Hormones and Behavior 55.3 (2009), pp. 375–89. issn: 1095-6867 (Electronic) 0018-506X (Linking). doi: 10.1016/j.yhbeh.2008.12.009.

[167] M. Rubinov and O. Sporns. “Complex network measures of brain connectivity: uses and interpretations”. In: NeuroImage 52.3 (2010), pp. 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003.

[168] T. Ruf and F. Geiser. “Daily torpor and hibernation in birds and mammals”. In: Biologica Reviews 90.3 (2015), pp. 891–926. doi: 10.1111/brv.12137.

[169] M. Ruse. Taking Darwin seriously. A naturalistic approach to philosophy. Prometheus, 1988.

[170] C. Sandi and J. Haller. “Stress and the social brain: behavioural effects and neurobiological mechanisms”. In: Nature Reviews Neuroscience 16.5 (2015), pp. 290–304. doi: 10.1038/nrn3918.

[171] N. Scafetta, D. Marchi, and B. J. West. “Understanding the complexity of human gait dynamics”. In: Chaos 19.2 (2009), p. 026108. issn: 1089-7682 (Electronic) 1054-1500 (Linking). doi: 10.1063/1.3143035. url: https://www.ncbi.nlm.nih.gov/pubmed/19566268.

[172] R. Scheines et al. “The TETRAD Project: Constraint Based Aids to Causal Model Specification”. In: Multivariate Behavioral Research 33.1 (1998), pp. 65–117. issn: 0027-3171. doi: 10.1207/s15327906mbr3301_3. url: https://doi.org/10.1207/s15327906mbr3301_3.

[173] J. Schmidhuber. “Deep learning in neural networks: An overview”. In: Neural Networks 61 (2015), pp. 85–117. issn: 0893-6080. doi: https://doi.org/10.1016/j.neunet.2014.09.003. url: http://www.sciencedirect.com/science/article/pii/S0893608014002135.

[174] L. Schwabe and O. T. Wolf. “Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action”. In: Behavioural Brain Research 219.2 (2011), pp. 321–328. doi: 10.1016/j.bbr.2010.12.038.

[175] L. Schwabe and O. T. Wolf. “Stress prompts habit behavior in humans”. In: Journal of Neuroscience 29.22 (2009), pp. 7191–7198. issn: 1529-2401 (Electronic) 0270-6474 (Linking). doi: 10.1523/JNEUROSCI.0979-09.2009. url: https://www.ncbi.nlm.nih.gov/pubmed/19494141.

[176] A. K. Seth and K. Friston. “Active interoceptive inference and the emotional brain”. In: Philosophical Transactions of the Royal Society B: Biological Sciences 371.1708 (2016), p. 20160007. doi: 10.1098/rstb.2016.0007.

[177] P. Shannon et al. “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. In: Genome Research 13.11 (2003), pp. 2498–2504. doi: 10.1101/gr.1239303.

[178] R. Shwartz-Ziv and N. Tishby. “Opening the black box of deep neural networks via information”. In: arXiv:1703.00810 [cs.LG] (2017). url: https://arxiv.org/abs/1703.00810.

[179] D. Silver et al. “Mastering the game of Go without human knowledge”. In: Nature 550 (2017), p. 354. doi: 10.1038/nature24270.

[180] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: arXiv:1409.1556 [cs.CV] (2014). url: https://arxiv.org/abs/1409.1556.
[181] B. F. Skinner. *The behavior of organisms: An experimental analysis*. B.F. Skinner Foundation, 1990. ISBN: 0996453903.

[182] R. Smith, T. Parr, and K. Friston. “Simulating emotions: An active inference model of emotional state inference and emotion concept learning”. In: *Frontiers in Psychology* 10 (2019), p. 2844. doi: 10.3389/fpsyg.2019.02844.

[183] R. Smith et al. “An active inference approach to modeling concept learning”. In: *bioRxiv* (2019), p. 633677. doi: 10.1101/633677.

[184] A. Solway et al. “Optimal behavioral hierarchy”. In: *PLoS Computational Biology* 10.8 (2014), e1003779. ISSN: 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1003779. url: https://www.ncbi.nlm.nih.gov/pubmed/25122479.

[185] A. Solway and M. M. Botvinick. “Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates”. In: *Psychological Review* 119.1 (2012), p. 120. doi: 10.1037/a0026435.

[186] C. Song, S. Havlin, and H. A. Makse. “Origins of fractality in the growth of complex networks”. In: *Nature Physics* 2.4 (2006), pp. 275–281. ISSN: 1745-2481. doi: 10.1038/nphys266. url: https://doi.org/10.1038/nphys266.

[187] C. Song, S. Havlin, and H. A. Makse. “Self-similarity of complex networks”. In: *Nature* 433.7024 (2005), pp. 392–395. ISSN: 1476-4687. doi: 10.1038/nature03248. url: https://doi.org/10.1038/nature03248.

[188] C. O. S. Sorzano, J. Vargas, and A. P. Montano. “A survey of dimensionality reduction techniques”. In: *arXiv:1403.2877 [stat.ML]* (2014). url: https://arxiv.org/abs/1403.2877.

[189] D. J. Spiegelhalter et al. “Bayesian measures of model complexity and fit”. In: *Journal of the Royal Statistical Society: Series B* 64.4 (2002), pp. 583–639. doi: 10.1111/1467-9868.00353.

[190] O. Sporns. “Network attributes for segregation and integration in the human brain”. In: *Current Opinion in Neurobiology* 23.2 (2013), pp. 162–171. ISSN: 0959-4388. doi: https://doi.org/10.1016/j.conb.2012.11.015. url: http://www.sciencedirect.com/science/article/pii/S0959438812001894.

[191] O. Sporns and J. D. Zwi. “The small world of the cerebral cortex”. In: *Neuroinformatics* 2.2 (2004), pp. 145–162. doi: 10.1385/NI:2:2:145.

[192] A. Staniloiu and H. Markowitsch. “P-246: The neuroimaging of dissociative disorders”. In: *European Psychiatry* 27 (2012), p. 1. doi: 10.1016/S0924-9338(12)74413-9.

[193] L. Stankov. “The structure among measures of personality, social attitudes, values, and social norms”. In: *Journal of Individual Differences* 28.4 (2007), pp. 240–251. doi: 10.1027/1614-0001.28.4.240.

[194] K. Starcke et al. “Does stress alter everyday moral decision-making?” In: *Psychoneuroendocrinology* 36.2 (2011), pp. 210–219. doi: 10.1016/j.psyneuen.2010.07.010.

[195] G. Storz and R. Hengge. *Bacterial stress responses*. American Society for Microbiology Press, 2010. ISBN: 1555816215.

[196] Y. Sun, F. Gomez, and J. Schmidhuber. “Planning to be surprised: Optimal Bayesian exploration in dynamic environments”. In: *Artificial General Intelligence. AGI 2011. Lecture Notes in Computer Science*. Springer, 2011, pp. 41–51. doi: 10.1007/978-3-642-22887-2_5.

[197] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, 2018. ISBN: 0262352702.

[198] S. Sutton. *Woodlice*. Elsevier, 2013. ISBN: 1483285782.

[199] N. Sünderhauf et al. “The limits and potentials of deep learning for robotics”. In: *International Journal of Robotics Research* 37.4-5 (2018), pp. 405–420. doi: 10.1177/0278364918770733.
I. Tagkopoulos, Y.-C. Liu, and S. Tavazoie. “Predictive behavior within microbial genetic networks”. In: *Science* 320.5881 (2008), pp. 1313–1317. doi: 10.1126/science.1154456.

J. R. Talevich *et al.*. “Toward a comprehensive taxonomy of human motives”. In: *PLoS ONE* 12.2 (2017), e0172279. issn: 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0172279. url: https://www.ncbi.nlm.nih.gov/pubmed/28231252.

J. B. Tenenbaum *et al.*. “How to grow a mind: Statistics, structure, and abstraction”. In: *Science* 331.6022 (2011), pp. 1279–1285. doi: 10.1126/science.1192788.

A. R. Todd *et al.*. “Anxious and egocentric: How specific emotions influence perspective taking”. In: *Journal of Experimental Psychology: General* 144.2 (2015), p. 374. doi: 10.1037/xge0000048.

U. Toelch and R. J. Dolan. “Informational and normative influences in conformity from a neurocomputational perspective”. In: *Trends in Cognitive Sciences* 19.10 (2015), pp. 579–589. doi: 10.1016/j.tics.2015.07.007.

G. Tononi, O. Sporns, and G. M. Edelman. “A measure for brain complexity: relating functional segregation and integration in the nervous system”. In: *Proceedings of the National Academy of Sciences of the United States of America* 91.11 (1994), pp. 5033–5037. issn: 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.91.11.5033. url: https://www.ncbi.nlm.nih.gov/pubmed/8197179.

D. Tran *et al.* “Learning spatiotemporal features with 3D convolutional networks”. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision*. 2015, pp. 4489–4497. doi: 10.1109/ICCV.2015.510.

H. Ursin and H. R. Eriksen. “Cognitive activation theory of stress (CATS)”. In: *Neuroscience & Biobehavioral Reviews* 34.6 (2010), pp. 877–881. doi: 10.1016/j.neubiorev.2009.03.001.

D. C. Van Essen *et al.* “The WU-Minn human connectome project: an overview”. In: *Neuroimage* 80 (2013), pp. 62–79. doi: 10.1016/j.neuroimage.2013.05.041.

J Van Oort *et al.* “How the brain connects in response to acute stress: A review at the human brain systems level”. In: *Neuroscience & Biobehavioral Reviews* 83 (2017), pp. 281–297. doi: 10.1016/j.neubiorev.2017.10.015.

J. Vasil *et al.* “A World unto Itself: Human Communication as Active Inference”. In: *Frontiers in Psychology* (2020). doi: 10.3389/fpsyg.2020.00417.

S. P. Veissièrre et al.* “Thinking through other minds: A variational approach to cognition and culture”. In: *Behavioral and Brain Sciences* (2019), pp. 1–97. doi: 10.1017/S0140525X19001213.

A. J. Veraart *et al.* “Recovery rates reflect distance to a tipping point in a living system”. In: *Nature* 481.7381 (2012), pp. 357–359. issn: 1476-4687. doi: 10.1038/nature10723. url: https://doi.org/10.1038/nature10723.

B. Von Dawans *et al.* “The social dimension of stress reactivity: acute stress increases prosocial behavior in humans”. In: *Psychological Science* 23.6 (2012), pp. 651–660. doi: 10.1177/0956797611431576.

C. S. Wallace and D. L. Dowe. “Minimum message length and Kolmogorov complexity”. In: *The Computer Journal* 42.4 (1999), pp. 270–283. doi: 10.1093/comjnl/42.4.270.

D. M. Walsh. *Organisms, agency, and evolution*. Cambridge University Press, 2015. isbn: 1107122104.

E. Watson, L. S. Yilmaz, and A. J. Wallhout. “Understanding metabolic regulation at a systems level: metabolite sensing, mathematical predictions, and model organisms”. In: *Annual Review of Genetics* 49 (2015), pp. 553–575. doi: 10.1146/annurev-genet-112414-055257.

D. J. Watts and S. H. Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *Nature* 393.6684 (1998), pp. 440–442. doi: 10.1038/30918.
[18] M. D. Wheelock et al. “Psychosocial stress reactivity is associated with decreased whole-brain network efficiency and increased amygdala centrality”. In: *Behavioral Neuroscience* 132.6 (2018), pp. 561–572. doi: 10.1037/bne0000276.

[19] J. C. Wingfield. “Control of behavioural strategies for capricious environments”. In: *Animal Behaviour* (2003), pp. 807–816. doi: 10.1006/anbe.2003.2298.

[20] J. C. Wingfield et al. “Ecological bases of hormone–behavior interactions: the “emergency life history stage””. In: *American Zoologist* 38.1 (1998), pp. 191–206. doi: 10.1093/icb/38.1.191.

[21] D. L. K. Yamins and J. J. DiCarlo. “Using goal-driven deep learning models to understand sensory cortex”. In: *Nature Neuroscience* 19 (2016), p. 356. doi: 10.1038/nn.4244. url: https://doi.org/10.1038/nn.4244.

[22] C. Yan and Y. He. “Driving and driven architectures of directed small-world human brain functional networks”. In: *PLoS ONE* 6.8 (2011), e23460. doi: 10.1371/journal.pone.0023460.

[23] R. Young. “A General Architecture for Robotics Systems: A Perception-Based Approach to Artificial Life”. In: *Artificial Life* 23.2 (2017), pp. 236–286. doi: 10.1162/ARTL_a_00229.

[24] F. F. Youssef et al. “Stress alters personal moral decision making”. In: *Psychoneuroendocrinology* 37.4 (2012), pp. 491–498. doi: 10.1016/j.psyneuen.2011.07.017.

[25] H. Yu and M. Gerstein. “Genomic analysis of the hierarchical structure of regulatory networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.40 (2006), pp. 14724–14731. doi: 10.1073/pnas.0508637103.

[26] J. Zhao et al. “Hierarchical modularity of nested bow-ties in metabolic networks”. In: *BMC Bioinformatics* 7 (2006), pp. 386–386. issn: 1471-2105. doi: 10.1186/1471-2105-7-386. url: https://www.ncbi.nlm.nih.gov/pubmed/16916470https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560398/.

[27] T. Zhao, R. Zhao, and M. Eskenazi. “Learning discourse-level diversity for neural dialog models using conditional variational autoencoders”. In: *arXiv:1703.10960 [cs.CL]* (2017). url: https://arxiv.org/abs/1703.10960.

[28] H. Zhu et al. “Increased functional segregation of brain network associated with symptomatology and sustained attention in chronic post-traumatic stress disorder”. In: *Journal of Affective Disorders* 247 (2019), pp. 183–191. issn: 0165-0327. doi: https://doi.org/10.1016/j.jad.2019.01.012. url: http://www.sciencedirect.com/science/article/pii/S016503271831303X.

[29] J. Zimmermann et al. “The latent structure of personality functioning: Investigating criterion a from the alternative model for personality disorders in DSM–5”. In: *Journal of Abnormal Psychology* 124.3 (2015), pp. 532–548. doi: 10.1037/abn0000059.

[30] O. Zinchenko and M. Arsalidou. “Brain responses to social norms: Meta-analyses of fMRI studies”. In: *Human Brain Mapping* 39.2 (2018), pp. 955–970. doi: 10.1002/hbm.23895.