Recent advances in the application of predictive coding and active inference models within clinical neuroscience

Ryan Smith, PhD[1], Paul Badcock, PhD[2,3,4], Karl J. Friston, MRCPsych[5]

[1]Laureate Institute for Brain Research, Tulsa, OK, USA

[2]Centre for Youth Mental Health, The University of Melbourne, Victoria, 3052, Australia

[3]Orygen, Victoria, 3052, Australia

[4]Melbourne School of Psychological Sciences, The University of Melbourne, Victoria, 3052, Australia

[5]Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, WC1N 3BG, UK

Corresponding author:

Ryan Smith

Laureate Institute for Brain Research

6655 S Yale Ave, Tulsa, OK 74136, USA

Email: rsmith@laureateinstitute.org

Counts (Review Article):

Figures: 3

Abstract: 235 words

Manuscript: 7,256 words

Journal Field:

General topics in psychiatry and related fields

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/pcn.13138

Abstract

Research in clinical neuroscience is founded on the idea that a better understanding of brain (dys)function will improve our ability to diagnose and treat neurological and psychiatric disorders. In recent years, neuroscience has converged on the notion that the brain is a 'prediction machine'—in that it actively predicts the sensory input that it will receive if one or another course of action is chosen. These predictions are used to select actions that will (most often, and in the long-run) maintain the body within the narrow range of physiological states consistent with survival. This insight has given rise to an area of clinical computational neuroscience research that focuses on characterizing neural circuit architectures that can accomplish these predictive functions, and on how the associated processes may break down or become aberrant within clinical conditions. Here, we provide a brief review of examples of recent work on the application of predictive processing models of brain function to study clinical (psychiatric) disorders, with the aim of highlighting current directions and their potential clinical utility. We offer examples of recent conceptual models, formal mathematical models, and applications of such models in empirical research in clinical populations, with a focus on making this material accessible to clinicians without expertise in computational neuroscience. In doing so, we aim to highlight the potential insights and opportunities that understanding the brain as a prediction machine may offer to clinical research and practice.

Keywords: Computational Neuroscience; Computational Psychiatry; Predictive Coding; Active Inference; Emotion

Introduction

Research in clinical neuroscience is founded on the premise that better understanding brain function – and by extension, dysfunction – will improve our ability to diagnose and treat disorders of the mind and brain. As the field has moved forward, it has become increasingly clear that, to understand the brain, it must first be characterized at multiple spatiotemporal scales and levels of description. We then have to understand how dynamics at each scale or level of description relate to (or emerge from) the dynamics of others. At one end of the spectrum, a large body of work has emerged over the last century on the micro-scale cellular and molecular functions of neurons and supporting brain cells; at the other, the last few decades have produced a growing body of work on large-scale human brain function using neuroimaging methods, which has uncovered regularities in the relationship between macro-scale regional brain activity and complex psychological processes. However, to fully link micro- and macro-scale functioning, and biological and psychological levels of description, it is widely recognized that we need to develop a better understanding of functional architectures at the intermediate meso-scale (see **Figure 1**).

The meso-scale corresponds to circuits of interconnected neurons with particular patterns of connectivity that allow for specific types of information processing or *computation*. For instance, one pattern of neural circuit connections may allow the brain to infer what objects are most likely giving rise to its current retinal input, whereas other patterns of connections may solve the problem of determining which action is more likely to generate preferred outcomes (1-3). Thus, meso-scale circuits afford a computational level of description, which offers a bridge between biology and psychology by allowing researchers to characterize how neural circuit dynamics produce psychological operations like

object recognition and decision-making (3-7). Unfortunately, the meso-scale is currently the most difficult to study directly with current neuroscience methods. However, a more recent body of work in computational neuroscience – and its clinically focused sister discipline of computational psychiatry – has emerged over the last several years to better address this problem (8-10).

The unique advantages offered by computational neuroscience and psychiatry follow from a specific type of mathematical modeling. This approach first identifies a problem that the brain can (or must) solve, such as object recognition or decision-making, and then works backward to identify the requisite computations – or algorithms – that the brain could use to solve this problem. Although many mathematical solutions may exist, only a subset will be biologically plausible – that is, only some algorithms could be plausibly accomplished by connecting neurons together in particular patterns. Once a set of biologically plausible algorithms has been identified, each algorithm in turn entails one or more testable hypotheses about how the meso-scale function and structure of a particular brain region (or connected set of regions) could implement that algorithm. Different algorithms and implementations typically have different costs and benefits, and, as such, they often make distinct predictions about the patterns of brain activity and behavior that should be observed under specific conditions. Such predictions can then be tested experimentally to provide evidence for one algorithm versus another and, in some cases, evidence for one neural implementation (meso-scale circuit structure) over another. Therefore, the two primary advantages of computational neuroscience include: 1) building mathematical models that allow researchers to simulate neurocomputational processes – to confirm how the brain *could* solve a particular problem; and 2) making empirical predictions for testing which neurocomputational processes the brain in fact *does* use to solve that problem.

There are many algorithms currently used for research in this field. In this article, we focus on a family of biologically plausible algorithms that have emerged from the idea that brain processes are *predictive*. That is, they share the idea that the brain is a type of 'prediction machine', which allows it to solve difficult problems in perception, decision-making, and skeletomotor/visceromotor control in an approximately optimal manner. In perception, one widely studied algorithm of this kind is 'predictive coding', which rests on the idea that a perceptual system continually predicts what it should observe next if its current perceptual beliefs are correct (5, 6, 11). Perceptual beliefs are then updated when sensory input differs from predicted input. In predictive coding, this difference affords the calculation of a 'prediction error', which can be used to guide belief updating and find the minimal change in perceptual beliefs necessary to minimize that error signal. Technically, this is called Bayesian belief updating or inference. This process can also be iterated hierarchically, so that higher brain circuits continuously update beliefs about what is represented in lower hierarchical levels that are closer to the sensory periphery – allowing different levels of abstraction (e.g., a belief that a white/round percept corresponds to a baseball; (1)). Different predictions and prediction errors can also be considered more or less reliable (e.g., visual prediction errors may be more reliable during the day vs. at night). As such, in predictive coding models the brain can also maintain (and update) estimates about the reliability (inverse variance or 'precision') of its various predictions and prediction errors – which modulate how easy or hard it is to change current beliefs when challenged by strong prediction errors (i.e., increasing or decreasing the 'weight' assigned to those prediction errors).

To give the reader a sense of the dynamics of predictive coding models, and how they might be applied to solve a clinically relevant psychological task, a very simple, 4-neuron network, and simulated neuronal

activity, is shown in **Figure 2A-E.** The details of the example are described in the figure legend; but, in short, the figure illustrates the simulated neuronal dynamics through which prediction error is minimized to arrive at an estimate or 'best guess' about whether a person is hostile or friendly. This is based on 1) the perceived curvature of their lips (e.g., flat curvature, most consistent with a neutral facial expression) and 2) prior expectations (i.e., initial predictions about friendliness prior to seeing the facial expression). The simulations show how prior expectations can bias perception to favor a 'hostile' interpretation, and how different relative precision-weightings can amplify or attenuate this effect.

Formally, and under certain simplifying assumptions, minimizing prediction error is equivalent to the minimization of a statistical quantity called *variational free energy*, which is often used to solve optimization problems in machine learning (12). In this context, free energy is a computationally tractable measure of the evidence that sensory input provides for the brain's internal (generative) model of the world, such that minimizing free energy maximizes model evidence. Free energy is simultaneously a measure of complexity minus accuracy. This means that, when the brain updates beliefs by minimizing (precision-weighted) prediction errors (i.e., minimizing free energy), it does so by finding the simplest (i.e., least complex), most parsimonious change in beliefs to account for new observations.

In decision-making and action selection, a related class of biologically plausible algorithms comes from the 'active inference' framework (1, 13-15). Active inference models go a step beyond predictive coding to emphasize that the brain does not simply predict sensory input passively. Instead, it predicts what it will observe if it chooses to act in one way or another (16). From this perspective, decision-making

involves predicting the outcomes one *would* observe *if* each of several actions (or action sequences, called 'policies') were chosen, and then selecting the actions expected to produce the observations that are most preferred and/or that will provide the most information. Upon reflection, this is quite intuitive. For example, if I am hungry, then I am less likely to stay still and more likely to go and find food—precisely because being hungry is inconsistent with my preferences and I expect eating food will produce my more preferred feelings of satiety. However, if I am not confident where to find the closest source of food, I will first choose actions to gather information – such as checking in the fridge before deciding to go to the store.

In active inference, preferences are formally modelled as a particular type of prior expectation (over sensory observations or 'outcomes'; often called 'prior preferences'), which means that non-preferred outcomes are 'surprising' in a technical sense (i.e., they deviate from those prior preferences). During decision-making, this type of expected 'surprise' (given one choice vs. another) is associated with what might be thought of as an expected 'preference prediction error' (i.e., the expected difference [technically the Kullback-Leibler divergence] between preferred outcomes and the outcomes expected if one were to choose a particular course of action – a quantity also sometimes referred to as 'risk'). Preferred outcomes are then achieved by choosing actions expected to minimize this type of prediction error. However, it is important to distinguish this preference-based notion of 'surprise' and prediction error from other constructs. First, prediction error is not equivalent to a conscious feeling of surprise or the conscious belief that something is surprising. Conscious feelings and beliefs are higher-level states that must themselves be inferred (17-19). By contrast, in active inference (and in predictive coding) the terms 'belief' and 'surprise' are technical terms that refer to probabilistic representations encoded by

neuronal activity (and connectivity) in meso-scale neural circuitry (e.g., canonical microcircuits; (5)). These neurocomputational processes occur at a sub-personal level outside of conscious awareness and therefore need not lead to conscious feelings of surprise. Second, 'preference' prediction errors (also sometimes called 'outcome prediction errors' in the active inference literature; e.g., see Figure 2 within (19)) are not the same as the 'state' prediction errors discussed above in relation to predictive coding. There are also state prediction errors in active inference (see **Figure 2F**), which – as in predictive coding – occur when model beliefs (about one's current state or situation) are updated with an unexpected new observation. In other words, state prediction errors drive belief updating about states of the world that are encountered, while preference prediction errors pertain to the (non)preferred outcomes expected if one chose different actions. In this sense, it is possible to be 'surprised' by an unanticipated outcome given one's prior beliefs about states (i.e., large state prediction error), while yet anticipating a preferred outcome (i.e., small expected preference prediction errors). This speaks to the difference between posterior beliefs about one's course of action (that are informed by state prediction errors) and prior beliefs before experiencing the consequences of action (that are informed by expected preference prediction errors).

Similar to state prediction errors in predictive coding, minimizing expected preference prediction error also corresponds to minimizing free energy—but in this case it is the free energy associated with the future outcomes that are *expected* under different actions (i.e., expected free energy). As with predictive coding, active inference comes equipped with a neural process theory that allows one to simulate patterns of neural activity that can be tested empirically (1, 14, 15, 20, 21). An example neural network implementation based on this process theory is depicted in **Figure 2F**. Although the neural

network architecture in this case differs somewhat from the simpler predictive coding example in **Figure 2A-E**, similar prediction-error minimizing dynamics emerge (example simulations using this architecture are shown in relation to a formal model of emotion in **Figure 3** below). In most cases, minimizing (expected) free energy in these theories is expected (on average, and in the long-run) to keep an organism within the narrow range of physiological and environmental states consistent with its survival (i.e., preferred states; (22)). However, it is important to highlight that organisms also plausibly inherit evolutionarily selected, adaptive prior expectations (including preferences), which can favor genetic propagation over individual survival (e.g., risky reproductive behavior or self-sacrifice for genetic relatives; (23)).

Similar to predictive coding, probabilistic beliefs in active inference models also have precisions that influence belief updating. However, in addition to the precision of sensory prediction errors and prior beliefs about states, there are a number of other variables that have precisions in active inference. For example, prior preferences can have different precisions, which can affect the degree to which behavior is driven by seeking reward vs. seeking information to resolve uncertainty. Beliefs about the probability (i.e., value) of different available sequences of actions ('policies') can also have different precisions, which can affect how strongly decision making is controlled by habits vs. explicit planning. Beliefs about distant future states can also be more or less precise (e.g., more or less confidence in predicting the way one's life will be in 6 months if choosing one or another course of action now), which can influence whether someone focuses more on short-term vs. long-term consequences when making decisions. More generally, because precision formally refers to the inverse dispersion (i.e., entropy) of a probability distribution, every belief has the attribute of precision. In predictive coding schemes these probabilistic

beliefs are over continuous variables (e.g., brightness, loudness) and precision corresponds to the inverse variance of those (typically Gaussian) distributions. In contrast, because decisions are categorical (i.e., a person can only choose one discrete option out of several available options), probabilistic beliefs in active inference models are over categorical (discrete) variables. In this case, precision can correspond to what are called 'inverse temperature' parameters that influence the shape of discrete probability distributions. For example, the beliefs (probabilities) over policies mentioned above are controlled by an 'expected precision' parameter of this type. Physiologically, these types of precisions can be associated with lateral inhibition and excitation-inhibition balance, whereas in predictive coding schemes precision can be associated with postsynaptic gain or excitability.

Although considerable progress has been made in building and testing computational models of simple problems in perception, cognition, and action selection; applications of predictive coding and active inference (as well as other related Bayesian) models to more complex clinical phenomena have thus far been limited. In this domain, the fields of clinical computational neuroscience and computational psychiatry are still largely at the stage of building viable conceptual models—and identifying or simulating associated mathematical models of complex clinical phenomena. Thorough empirical tests of these models are largely outstanding, although emerging empirical work (discussed below) appears promising.

Illustrative examples

In what follows, we offer a number of illustrative examples of conceptual and formal mathematical models in computational neuroscience and psychiatry – with the aim of 1) illustrating potential empirical

applications, and 2) conveying the methodological resources available that can be applied to other cases. We then describe examples of recent empirical work that has applied such models to clinical questions. These examples should not be seen as an exhaustive treatment of extant literature—they just exemplify ways in which understanding the brain as a 'prediction machine' can be used in psychiatry. It should also be noted that we do not address other areas of research in computational psychiatry, e.g., reinforcement learning. The primary focus of reinforcement learning is to account for choice behavior in terms of rewards. This area of research has been fruitful, but it has focused less on modelling the prediction-based meso-scale brain processes we consider here (i.e., with some notable exceptions related to dopamine and reward prediction error signals, and their interactions with corticostriatal circuits; e.g., see (7, 24, 25)).

Conceptual models

There is a growing body of theoretical work that affords potential clinical insights and new hypotheses for psychiatric disorders. This work takes the general approach of synthesizing previous empirical findings and proposing ways in which computational concepts could help unify such findings under a single parsimonious theory. We consider these contributions to be 'conceptual' because, although they qualitatively describe ways in which the predictive neurocomputational dynamics illustrated in **Figure 2** may shed light on clinical phenomena, they do not offer quantitative mathematical models that afford precise simulations of proposed mechanisms.

As one prominent example, depression has been the topic of a number of computational proposals. One recent model pertains to a proposed computational basis for mood (26). The central idea here is that

neural systems not only maintain estimates in the confidence or precision of their prior beliefs about outcomes, but also estimates of confidence in their beliefs about that precision (i.e., the degree to which the unknowns are known). For example, if an individual expects uncertain, unpredictable outcomes (low precision beliefs), but makes *those* predictions confidently (high confidence in high uncertainty), this will result in a chronic, self-maintaining negative emotional state that, in extreme cases, becomes resistant to change (as in depressive disorders). Neurobiologically, the set-points of neuromodulatory mechanisms are suggested to encode these—statistically higher order—confidence estimates, potentially accounting for the neuromodulatory abnormalities observed in depression (and the mechanisms of antidepressant drugs in targeting neuromodulatory systems; (27, 28)). In neural process theories (**Figure 2**), these neuromodulators would have the effect of dynamically tuning the strengths of synapses that encode various types of precisions (e.g., the estimated reliability of sensory signals, prior expectations, expected action outcomes, etc.).

This model represents a nice example of how (sub-personal) beliefs, computations, and meso-scale neuronal processes can be understood in terms of each other. That is, the precision corresponds to the statistical certainty of a probabilistic representation, where this certainty is encoded by the patterns of synaptic efficacy or gain (and resulting excitability) within a neuronal population (see (3, 15)). The required computation is then to optimize this encoding of uncertainty, by adjusting synaptic connection strengths through neuromodulatory mechanisms (e.g., serotonergic, or dopaminergic neurotransmission). In this setting, 'optimization' means that neuronal dynamics will converge onto an internal estimate (a 'best guess' based on past experience) that minimizes free energy. In the case of psychopathology, early adversity (and perhaps genetic/epigenetic vulnerability factors) could lead an

individual to converge onto poor or otherwise maladaptive estimates of uncertainty, as suggested by the model described above.

Another predictive processing perspective has been offered by Barrett and colleagues (29), who focus on the role of interoceptive predictive processing and the ways in which interoceptive predictions become dysfunctional and result in a depressive phenotype. They propose that depression is due to one or more of the following: 1) an internal model that is inefficient at managing energy regulation (e.g., due to overly precise expectations for large metabolic demand), 2) imprecise interoceptive signals from the body, creating difficulty in effective allostasis (i.e., difficulty adaptively generating anticipatory changes in visceral states due to expected future demands), and 3) maladaptive internal estimates of the precision of afferent interoceptive signals (e.g., due to neuromodulatory system dysfunction). Each of these potential breakdowns would render the brain insensitive to updating its beliefs about the body, and lead to difficulty keeping the body within homeostatic ranges (i.e., 'preferred states' within active inference models). Under the assumption (put forward by the authors) that pleasant/unpleasant feelings convey interoceptive information about the moment-to-moment energy conditions (i.e., immunological, inflammatory, and physiological states) of the body, this could produce pervasive negative affect (and sickness behaviors that reduce energy expenditure and promote fatigue), and lead to several motivational and neurovegetative symptoms of depression. In previous work, these authors have also proposed an explicit meso-scale scheme in which cortical columns within agranular cortical regions within the insula and anterior cingulate convey interoceptive prediction signals, while granular cortical regions (e.g., posterior insula) generate interoceptive prediction error signals that update beliefs

about the energy conditions of the body (30) – offering additional predictions that could be tested empirically in future neuroimaging work.

A further conceptual model of depression has been proposed by Badcock and colleagues (31), which highlights how computational processes can interact with the various psychosocial processes that are implicated in depression. They propose an evolutionary systems theory of depressive mood states, which combines active inference with insights—drawn from psychology, psychiatry, and neuroscience—on the role of social contexts in depressive phenomena. Under this model, normative levels of depressed mood can reflect an adaptive, risk-averse strategy that reduces uncertainty in interpersonal contexts when sensory cues evince an increased likelihood of unexpected or negative social outcomes (e.g., rejection or loss). The depressive response thus functions as a 'better safe than sorry' strategy that minimizes interpersonal interactions with unpredictable or non-preferred expected outcomes. It achieves this function by inducing changes in perception (e.g., a heightened sensitivity to social information); suppressing confident reward-approach behaviors (e.g., anhedonia); and generating signaling behaviors that either garner support (e.g., reassurance seeking) or defuse conflict (e.g., submissive behaviors). The authors suggest that, neurobiologically, depressed mood states are characterized by an increase in the precision of (bottom-up) social prediction errors (i.e., prediction errors conveyed to neurons encoding social information), which facilitates perceptual inference and learning about the social world by increasing the influence of ascending prediction errors on belief updating. This idea is consistent with the active inference literature on emotion, which suggests that negatively valenced states increase the learning rate of the causes of sensory stimuli by increasing the precision of incoming prediction errors (see (32)). In depression, these amplified prediction errors are

postulated to be selective to interpretations of social stimuli, which most plausibly occur at higher levels in a hierarchical model (e.g., dorsomedial prefrontal cortices; (33-37)). This increase in precision amplifies an individual's sensitivity or attention to interpersonal cues, while suppressing confidence in (top-down) social predictions (and thus confidence in associated social behaviors). Symptomatically, the authors suggest that this would produce a suspension of goal-directed behavior (e.g., anhedonia and social withdrawal), rumination about negative self-other relations, and an attentional bias toward (aversive) interpersonal cues during social inference. Here, it is especially important to stress the role of attention. In this instance, attention can be construed as affording precision to neuronal signals conveying various Bayesian beliefs or prediction errors (38-43); such that they have greater influence on belief updating than other levels of the cortical hierarchy. This fits comfortably with the role of neuromodulators in mediating attentional gain at the synaptic level (3).

In most circumstances, the suggestion is that the depressive response functions adaptively by attracting interpersonal support and reducing social uncertainty, through both risk-averse interpersonal behaviors and faster belief updating in the presence of unexpected outcomes. However, psychopathology can emerge when there are ongoing discrepancies between actual and preferred social outcomes over time (i.e., chronic preference prediction errors). Given the increased precision-weighting assigned to social prediction errors, such chronic social stress can in turn engender precise higher-order expectations that social rewards are unlikely (e.g., pessimism, low self-worth), which perpetuate risk-averse depressive behaviors (e.g., social withdrawal) and ultimately lead to disorder (e.g., learned helplessness); also see (44, 45)). As we discuss below, these depressive beliefs can become increasingly entrenched and self-maintaining, producing overly precise prior expectations about aversive (social) outcomes that are

resistant to change. In other words, depressive disorders can be described as a maladaptive pattern of dysregulated defenses: when depressive changes fail to resolve social stress, the individual is at risk of entering a self-perpetuating, dysregulated state, leading to chronic illness behaviors that fail to respond to any improvements in the social domain. Vulnerability to psychopathology is facilitated by early exposure to social stress (e.g., parental abuse or neglect), which promotes prior beliefs that social outcomes are uncontrollable and heightens the sensitivity of stress response systems to interpersonal stressors (e.g., inflammatory immune responses; see (46)). Notably, the model described here also lends itself to empirical scrutiny. For example, a simple way to test this hypothesis would be to use neuroimaging or electrophysiological methods that capture prediction error suppression (e.g., trial-by-trial fluctuations in P300 amplitudes) in depressed versus non-depressed participants when presented with unpredictable social stimuli. The model also has implications for prevention and intervention efforts by underscoring the importance of strategies that aim to improve social environments or resolve interpersonal stress (e.g., interpersonal psychotherapy).

Note that these are simply illustrative examples of recent conceptual models. As we touched upon above, other authors have proposed that depression is maintained by overly precise prior expectations for depressive schemas, such as strong expectations for one's own worthlessness or that the world is uncontrollable (47). The suggestion here is that these prior expectations bias attention to schema-consistent information and also bias the interpretation of sensory input in a schema-consistent manner. This can then drive chronic stress, heightened inflammation, and avoidance behaviors, each of which furnish future observations that selectively support and maintain depressive schemas in a vicious positive feedback loop (also see (45)). Further examples include recent conceptual models of autism (48-

50) and schizophrenia (51, 52). In autism it is suggested that, from childhood, low-level sensory prediction errors are over-weighted generally, leading to 1) repetitive behaviors that reduce low-level prediction error, and 2) a reduced ability to learn abstract and complex regularities in higher levels of a hierarchical model (i.e., because all prediction-errors are inappropriately suppressed before reaching these higher levels). As social regularities are among the most complex to learn, social cognition thus remains poorly developed. Interestingly, accounts of schizophrenia suggest that a similar phenomenon occurs, in which low-level prediction errors in some modalities are over-weighted, but with adult onset. This then leads the brain to treat random aspects of sensory observations as 'suspicious coincidences' in need of explanation, driving the development of overly complex models of the world with stranger and stranger (delusional) beliefs over time. For example, when proprioceptive prediction errors (that carry information about bodily movement) are over-weighted, this can lead to delusions about agency (e.g., that unexplained proprioceptive prediction errors indicate that one's body is being controlled by others). In contrast, it appears that auditory prediction errors are under-weighted in psychosis, creating increased vulnerability to auditory illusions. Although further conceptual models have been proposed for other clinical phenomena, the proposals on depression covered in detail here are representative of the general explanatory strategies used in other cases.

Formal models

Moving beyond conceptual proposals, a few formal (i.e., quantitative mathematical) predictive processing models have also been presented for a range of clinical phenomena; unlike conceptual models, these formal models afford explicit simulations of proposed neurocomputational mechanisms.

In the case of depression, for example, Stephan et al. (53) presented a mathematical model to capture homeostatic and allostatic regulation of visceral states, and how metacognitive estimates of the efficacy of allostasis could promote a depression-like state of fatigue. In their proposal, homeostatic set points are modeled as fixed probability distributions specifying predicted ranges (means and variances) of a given physiological variable (e.g., blood glucose levels, blood osmolality levels, circulating hormone or inflammatory cytokine levels, etc.); deviations from homeostatic ranges lead to the generation of predictions errors (with respect to those set points), which can then be minimized through closed-loop control processes in an interoceptive reflex arc (i.e., this is analogous to minimizing 'preference prediction errors' in active inference models, where homeostatic set points represent preferred states). Allostasis involves forecasting future deviations from homeostasis and adjusting physiological variables in advance to prevent these deviations. In their model, the prior expectations for homeostatic ranges can be dynamically adjusted by higher-level predictions. For example, if a cue indicated that blood glucose levels were about to drop, the predicted mean of the distribution could be temporarily shifted upward—so that blood glucose levels would elevate before the anticipated drop, allowing them to always remain within acceptable values. Alternatively, if a cue indicated that blood glucose levels were about to change in an unexpected direction, the predicted precision of the distribution could be tightened, endowing prediction errors with a greater influence on belief updating—and a faster return to homeostatic ranges. They then discuss how, if attempts at allostasis repeatedly fail, sustained dyshomeostasis could lead to higher-level inferences favoring states of subjective fatigue associated with depression. Specifically, they suggest that such states of depressive fatigue, and associated sickness behaviors, could emerge as strategies for dealing with conditions in which the brain continuously

predicts that allostatic regulation will be ineffective. (To be clear, we use blood glucose levels here only as an example physiological variable for illustration. The authors focus on dyshomeostasis generally – consistent with the conceptual proposals regarding metabolic regulation in depression reviewed above.)

Crucially, the quantitative prediction and prediction-error dynamics that can be simulated using this model speak to the possibility of testing which regions of the brain show patterns of activity consistent with those simulated dynamics, and whether this differs in healthy vs. depressed individuals. The authors hypothesize that a number of subcortical and cortical regions (e.g., brainstem, amygdala, insula, anterior cingulate) may implement different levels of homeostatic and allostatic control, and that dorsal prefrontal regions may be involved in estimating allostatic efficacy. Testing these predictions will be an important direction for future empirical work.

As a further, more in-depth example of recent formal modelling, another series of papers has presented a formal active inference model of emotional state inference and emotional awareness (depicted in **Figure 3**), and used this model to simulate clinical phenomena arising from poor early learning and biased attentional mechanisms (19, 54); also see (18). This modelling effort aimed to explain the common clinical observation that some individuals appear to have low awareness of their own emotions and can misinfer that emotion-related sensations are signs of medical illness (55-57); it is also inspired by the observation that psychotherapeutic interventions that aim to improve emotional awareness have shown efficacy (58, 59). To simulate these phenomena, the model included examples of distinct internal states that could be inferred, including emotional state categories (sadness, panic) and somatic state categories (sickness, heart attack). The simulated decision-making 'agent' began with prior expectations

about its own internal state, and these expectation were then updated (i.e., a new internal state was inferred) based on examples of relevant lower-level beliefs about current experience, including: valence (in this case, neutral/negative), arousal (high/low), motivation (approach/avoid), and an interpretation of the current situation (neutral, socially threatening, or physically threatening; e.g., as in cases where one is ignored at a party, in a crowded space, or feeling chest pain, etc.). To allow for biases in attention, the agent could not access all of this lower-level information at the same time. It instead had to selectively attend to one or more of these pieces of information sequentially (e.g., choose to attend to valence, then arousal, etc.) until it became confident in its internal state. At this point it would choose to self-report its internal state beliefs (e.g., whether it was experiencing sadness, panic, sickness, or a heart attack, or if it simply felt neutral, bad, or good). After making a report, the agent also received 'social feedback' that was either positive (if its report was 'correct''; e.g., matched cultural expectation) or negative (if it was incorrect). Positive social feedback was preferred. Thus, on each 'trial', the agent was presented with a newly generated 'affective response' (i.e., a pattern of valence, arousal, and motivation in a perceived context), and chose what to attend to and report in hopes of receiving the preferred positive feedback (i.e., to minimize preference prediction error). Each affective response pattern was generated by a 'true model', which held the different (probabilistic) patterns of lower-level information most consistent with each possible emotional/somatic state (e.g., 'sadness' best matched a pattern of negative valence, low arousal, and avoidance motivation in the context of social threat [such as being socially rejected]).

Aside from this inference process, a variant of the model also implemented emotion concept learning. Specifically, the simulated agent could learn emotion categories (e.g., sadness, fear, anger, and

happiness) that predicted different patterns in experience through the social feedback mentioned above (e.g., as in early development, when caregivers mirror and label a child's reactions with emotion terms (60-62)). Mathematically, this learning process (i.e., a type of concept acquisition) was implemented through learning probability distributions, via a coincidence detection mechanism resembling Hebbian synaptic plasticity (14, 63, 64). This learning mechanism estimates how often different internally represented states (e.g., sadness) are expected to generate particular thoughts, feelings, and sensations (e.g., unpleasant feelings, high arousal sensations, and avoidance drives while perceiving a predator). The formal basis of this learning mechanism can be intuitively thought of as simply counting coincidences between states and observations – that is, increasing the expected probability (i.e., synaptic connection strength) of a particular observation, given a particular state, each time one makes that observation when (they believe) they are in that state – similar to long-term potentiation (LTP) and long-term depression (LTD) types of synaptic learning mechanisms widely studied in cellular/molecular neuroscience (14, 64).

These simulations are based on the 'three-process model' of emotional awareness (57, 65-70), and supporting evidence (56, 71-79), which emphasizes that the mechanisms that generate affective responses can be separated from the processes that infer the meaning of those responses. It is also based on elements of 'constructivist' theories that stress the culture- and context-specific nature of emotion conceptualization processes (80). Based on the evidence supporting these perspectives, the broad theoretical assumptions of the simulations are that individuals generate multimodal affective responses in a flexible (i.e., domain-general) and context-sensitive manner, based on the predicted metabolic, cognitive, and behavioral demands of a perceived (or remembered/imagined) situation.

These predicted demands are in turn based on a fast, largely unconscious evaluation of that situation (e.g., concerning whether it is safe/threatening, goal-congruent/incongruent, consistent/inconsistent with norms/values, among other dimensions) and available actions. Generated responses include the types of changes in motivation, physiological arousal levels, and felt valence mentioned above. Once these bodily sensations and drives have been generated and perceived, an individual must infer their meaning. Crucially, because mappings between emotion concepts and response patterns are learned (within a culture and with context-specificity) and probabilistic (e.g., sadness may typically be low-arousal but could be high-arousal), recognizing emotions can be a difficult inference problem – that can sometimes lead to poor understanding and awareness of one's own emotions and their causes.

Note that we have therefore moved from Bayesian beliefs to a computational description of feelings and emotions; namely, Bayesian beliefs that provide the best explanation for both exteroceptive and interoceptive signals. For example, a belief such as 'I am anxious' provides the most parsimonious account of perceived threat and autonomic arousal—that is itself influenced (in part) by the predictions ensuing from an expectation that one will feel anxious. This may sound rather abstract; however, this kind of belief updating is now possible to simulate in a neurobiologically plausible way and can, in principle, be used to produce symptoms of emotional pathology *in silico*. For example, the top right portion of **Figure 3** depicts simulated neuronal firing rates and local field potentials during this emotional state inference process, based on the neural process theory depicted in the right portion of **Figure 2**. Empirically, one can then take these meso-scale simulations and use neuroimaging (or other neuroscience methods) to identify which brain regions most plausibly implement those meso-scale computations (e.g., by finding brain regions that show the predicted patterns of activity in simulations).

Simulations in this model illustrated at least 7 different mechanisms that could promote low emotional awareness. Because emotional awareness (i.e., the ability to recognize and understand one's own emotions) plays a key role in many psychiatric disorders (55, 57), and may act as either a vulnerability or maintenance factor, understanding such mechanisms could represent an important step in being able to identify which mechanisms are operative in different individuals and how they might be targeted on an individual basis within therapy. For instance, one mechanism involved having overly precise prior expectations for somatic threats, which led the simulated agent to somatize (i.e., mistake sadness-related sensations as signs of sickness or mistake panic-related sensations as signs of a heart attack; e.g., as in high anxiety sensitivity (81)). In another example, if the agent's emotion concepts made highly imprecise predictions (e.g., as in poor emotion concept acquisition due to social neglect in childhood; (82, 83)), it remained unconfident and only reported feeling good or bad (i.e., it was unable to differentiate different types of unpleasant emotions; for clinical relevance, see (84, 85)).

A third mechanism involved selective attention biases, which led the agent to selectively ignore either bodily signals or contextual cues (i.e., where such attention biases could have been reinforced in childhood and prevented emotion concept learning; e.g., hypervigilant external attention, due to consistently high levels of threat or environmental unpredictability (86, 87)). Importantly, these mechanisms produced similar behavioral phenotypes (i.e., self-report patterns), but might be best targeted by different psychotherapeutic interventions. For example, if low emotional awareness is due to an individual's emotion concept representations generating imprecise predictions, this would suggest the need for psychoeducation interventions to help an individual mindfully attend to emotion and develop more precise emotion concept knowledge (e.g., mindfulness-based or emotion-focused

therapies; (88-90)). In contrast, if an individual instead has overly precise prior expectations for somatic threats, interventions that focus on attenuating those strong expectations (e.g., exposure, cognitive restructuring; (91)) may be more appropriate. Finally, the model also affords quantitative simulations of, and makes distinct predictions about, the neural responses and reaction time differences that would be observed if different mechanisms were involved. For example, faster reaction times ('jumping to conclusions') are predicted with overly precise prior expectations (e.g., for somatic threats). As another example, weaker neural responses to affective stimuli are predicted in individuals with less precise emotion concept representations, due to weaker changes in beliefs about emotional states upon the generation of a new affective response. Thus, these types of mathematical simulations afford important predictions to be tested in future experimental work that could identify which brain regions implement the meso-scale neurocomputational processes captured in these simulations.

We stress here again that these are simply illustrative examples of relevant simulation work in this area and are not meant to be exhaustive. For example, other recent work in active inference has mathematically simulated the factors that influence patients' decisions to adhere to antidepressant medications (92), and the neurocomputational mechanisms of change during cognitive and behavioral therapies (93). Yet other studies have simulated predictive processing dynamics in neural network and robot models and reproduced patterns of perception and behavior reminiscent of both schizophrenia (94) and autism (95, 96). However, the detailed examples above – which have selectively focused on depression and emotion – should offer the reader a foundational sense of current directions in this field.

Empirical studies

Recent neuroimaging studies have begun to support the hypothesis that the brain engages in predictive processing, both in perception and decision-making. As might be intuited by the foregoing theoretical review, many of these studies rest upon the relationship between aberrant precision-weighting and the consequent neuromodulation of brain responses—or connectivity: please see references in (97). These studies have been largely conducted with healthy participants. For example, a few studies have required participants to perform probabilistic perception and inference tasks, fit computational models to task behavior (98), and then generated simulated sequences of precision-weighted prediction errors (99, 100). Using neuroimaging, these studies have found patterns of brain activation that closely match the simulated pattern of prediction errors. The observed patterns of brain activation support the roles of dopamine (midbrain activity) and acetylcholine (septum and basal forebrain) in encoding and precision-weighting distinct types of prediction errors, associated with sensory predictions and uncertainty estimates, respectively. Another study (20) has also fit decision behavior to an active inference model and found evidence that activity within brain regions implicated in either sending or receiving dopaminergic signals closely tracks the difference in expected free energy before and after a new observation. This difference updates confidence estimates in the ability to select optimal actions (i.e., the expected certainty in achieving preferred outcomes), encoded by an 'inverse temperature parameter' reflecting the expected precision of beliefs about the expected free energy of policies (i.e., possible sequences of actions). This parameter subsequently influences how goal-directed an individual's actions are; imprecise beliefs about policies (i.e., low expected precision) will give way to more random or habit-driven behavior. Note, however, that this type of (policy-related) precision is distinct from the types of precision discussed above in predictive coding, which instead correspond to

beliefs about the reliability (inverse variance) of prior expectations and sensory prediction errors – and modulate how strongly beliefs are updated by sensory prediction errors (i.e., based on those beliefs about their reliability).

There are also some very recent predictive processing-based behavioral and neuroimaging studies in psychiatric populations (101, 102), which have provided evidence, for example, of heightened prior beliefs that the environment is unpredictable in schizophrenia patients and in those with high risk for psychosis (associated with activation patterns within prefrontal and insula regions). While such neuroimaging studies are quite limited, purely behavioral studies have also found support for computational differences in patient populations. For example, one study asked individuals with and without auditory hallucinations to indicate when they heard a tone each time a light was presented, while (unbeknownst to participants) the probability of the tone being presented changed over time (103). Computational modeling revealed that individuals with hallucinations overestimated the precision of auditory predictions, leading them to report false auditory percepts more often than non-hallucinators (for related work linking aberrant prior expectations and precision in psychosis, see (51, 104)). Interestingly, studies combining pupillometry and probabilistic learning tasks have shown a distinct pattern in autism, where similar modeling methods suggest that the precision of sensory predictions is under-weighted (48, 49, 105).

Lastly, a few recent studies have fit active inference models to data in transdiagnostic patient populations from tasks involving exploratory decision-making, approach-avoidance conflict, and interoceptive awareness. One study found lower learning rates for losses vs. wins, and lower precision in

action selection, within individuals with substance use disorders relative to healthy controls (106), suggesting a pattern in which actions that lead to negative outcomes fail to influence future decisions – which could help explain continued substance use despite negative life consequences. A second study found that, during approach-avoidance conflict, individuals with depression, anxiety, and/or substance use disorders each showed evidence of lower expected policy precision (greater uncertainty in decision-making; associated with the γ term in **Figure 2F**) in comparison to healthy controls (107). However, the clinical populations did not show heightened aversion to negative stimuli, suggesting that decision uncertainty may represent a more promising target for intervention/treatment. Finally, an active inference model of an interoceptive awareness task recently found evidence that individuals with depression, anxiety, substance use, and/or eating disorders had lower interoceptive sensory precision than healthy controls, but no difference in prior expectations (108). This finding was specific to an interoception-enhancing breath-hold manipulation, suggesting a transdiagnostic inability to update sensory precision estimates when there are changes in afferent signals from the body. This might help to explain visceral dysregulation and poor emotional awareness in these clinical populations. Such findings are promising, but will need to be replicated and extended in future work before being afforded high confidence.

Conclusion

We have reviewed illustrative examples of how modelling the brain as a prediction machine has begun to yield potential clinical insights. Initially, this involves proposing conceptual models of how a computational framework may offer explanatory power. Subsequently, formal computational models

can be written down, affording quantitative simulations and precise predictions. Finally, those models can be translated into behavioral tasks that can be carried out during neuroimaging. Models can then be fit to behavior, allowing both: 1) identification of the neural correlates of model parameters, and 2) identification of differences in model parameters (and associated neural responses) between individuals, which could offer either diagnostic or prognostic information and inform treatment selection—sometimes called computational phenotyping. This therefore allows us to move from meso-scale neurocomputational processes that are very difficult to study directly (e.g., through single-neuron recordings) all the way through to phenotyping individuals within clinical populations with heterogenous underlying mechanisms. Further advances in the application of formal models of the predictive brain will require close collaboration between clinicians, patients, and computational researchers to generate and formalize models of distinct clinical phenomena, and to design informative behavioral tasks that, once validated, patients could perform in the clinic to help mental health professionals better understand and treat their patients.

Acknowledgments. RS is funded by the William K. Warren Foundation. There is no other funding to report.

Disclosure Statement. The authors declare no conflict of interest.

Author Contributions. RS conceptualized and wrote the first draft of the manuscript and figures. PB aided in generating figures and wrote/edited sections of the manuscript. KJF contributed to conceptualization and also wrote/edited sections of the manuscript.

References

1. Friston K, Parr T, de Vries B. The graphical brain: Belief propagation and active inference. Network Neuroscience. 2017;1:381-414.
2. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 2004;27(12):712-9.
3. Parr T, Friston K. The Anatomy of Inference: Generative Models and Brain Structure. Frontiers in Computational Neuroscience. 2018;12:90.
4. Marr D. Vision. 1982.
5. Bastos A, Usrey W, Adams R, Mangun G, Fries P, Friston K. Canonical microcircuits for predictive coding. Neuron. 2012;76:695-711.
6. Friston K. A theory of cortical responses. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2005;360:815-36.
7. Frank M. Computational models of motivated action selection in corticostriatal circuits. Current Opinion in Neurobiology. 2011;21:381-6.
8. Friston K, Stephan K, Montague R, Dolan R. Computational psychiatry: the brain as a phantastic organ. The lancet Psychiatry. 2014;1:148-58.
9. Huys Q, Maia T, Frank M. Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience. 2016;19:404-13.
10. Montague P, Dolan R, Friston K, Dayan P. Computational psychiatry. Trends in Cognitive Sciences. 2012;16:72-80.
11. Kiebel S, Daunizeau J, Friston K. A Hierarchy of Time-Scales and the Brain. PLoS Computational Biology. 2008;4:e1000209.
12. Friston K. The free-energy principle: a unified brain theory? Nature reviews Neuroscience. 2010;11:127-38.
13. Da Costa L, Parr T, Sajid N, Veselic S, Neacsu V, Friston K. ACTIVE INFERENCE ON DISCRETE STATE-SPACES – A SYNTHESIS. arXiv. 2020:2001.07203v2 [q-bio.NC]
14. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, O Doherty J, Pezzulo G. Active inference and learning. Neuroscience and biobehavioral reviews. 2016;68:862-79.
15. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G. Active Inference: A Process Theory. Neural Computation. 2017;29:1-49.
16. Kaplan R, Friston KJ. Planning and navigation as active inference. Biol Cybern. 2018;112(4):323-43.
17. Whyte C, Smith R. The Predictive Global Neuronal Workspace: A Formal Active Inference Model of Visual Consciousness. bioRxiv. 2020:2020.02.11.944611.
18. Hesp C, Smith R, Allen M, Friston K, Ramstead M. Deeply Felt Affect: The Emergence of Valence in Deep Active Inference. PsyArXiv. 2019.
19. Smith R, Lane RD, Parr T, Friston KJ. Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance. Neurosci Biobehav Rev. 2019;107:473-91.
20. Schwartenbeck P, FitzGerald T, Mathys C, Dolan R, Friston K. The Dopaminergic Midbrain Encodes the Expected Certainty about Desired Outcomes. Cerebral Cortex. 2015;25:3434-45.

21. Schwartenbeck P, Friston K. Computational Phenotyping in Psychiatry: A Worked Example. eNeuro. 2016;3:ENEURO.0049-16.2016.
22. Badcock PB, Friston KJ, Ramstead MJD. The hierarchically mechanistic mind: A free-energy formulation of the human psyche. Phys Life Rev. 2019;31:104-21.
23. Badcock PB, Friston KJ, Ramstead MJD, Ploeger A, Hohwy J. The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behavior. Cogn Affect Behav Neurosci. 2019;19(6):1319-51.
24. Silvetti M, Alexander W, Verguts T, Brown J. From conflict management to reward-based decision making: actors and critics in primate medial frontal cortex. Neuroscience and biobehavioral reviews. 2014;46 Pt 1:44-57.
25. Schultz W. Dopamine reward prediction error coding. Dialogues Clin Neurosci. 2016;18(1):23-32.
26. Clark J, Watson S, Friston K. What is mood? A computational perspective. Psychological Medicine. 2018:1-8.
27. Cowen P, Browning M. What has serotonin to do with depression? World Psychiatry. 2015;14:158-60.
28. Hieronymus F, Nilsson S, Eriksson E. A mega-analysis of fixed-dose trials reveals dose-dependency and a rapid onset of action for the antidepressant effect of three selective serotonin reuptake inhibitors. Transl Psychiatry. 2016;6(6):e834.
29. Barrett LF, Quigley KS, Hamilton P. An active inference theory of allostasis and interoception in depression. Philos Trans R Soc Lond B Biol Sci. 2016;371(1708).
30. Barrett L, Simmons W. Interoceptive predictions in the brain. Nature reviews Neuroscience. 2015;16:419-29.
31. Badcock PB, Davey CG, Whittle S, Allen NB, Friston KJ. The Depressed Brain: An Evolutionary Systems Theory. Trends Cogn Sci. 2017;21(3):182-94.
32. Joffily M, Coricelli G. Emotional Valence and the Free-Energy Principle. PLoS Computational Biology. 2013;9:e1003094.
33. Adolphs R. The social brain: neural basis of social knowledge. Annual review of psychology. 2009;60:693-716.
34. Amodio DM, Frith CD. Meeting of minds: the medial frontal cortex and social cognition. Nature Reviews Neuroscience. 2006;7:268-77.
35. Meyer M, Lieberman M. Social working memory: Neurocognitive networks and directions for future research. Frontiers in Psychology. 2012;3:1-11.
36. Meyer M, Spunt R, Berkman E, Taylor S, Lieberman M. Evidence for social working memory from a parametric functional MRI study. Proceedings of the National Academy of Sciences of the United States of America. 2012;109:1883-8.
37. Meyer M, Taylor S, Lieberman M. Social working memory and its distinctive link to social cognitive ability: an fMRI study. Social cognitive and affective neuroscience. 2015;10:nsv065-.
38. Parr T, Friston K. Working memory, attention, and salience in active inference. Scientific Reports. 2017;7:14678.
39. Feldman H, Friston K. Attention, uncertainty, and free-energy. Frontiers in human neuroscience. 2010;4:215.

40. Parr T, Friston K. Uncertainty, epistemics and active inference. Journal of the Royal Society, Interface. 2017;14.
41. Mirza MB, Adams RA, Friston K, Parr T. Introducing a Bayesian model of selective attention based on active inference. Sci Rep. 2019;9(1):13915.
42. Parr T, Friston KJ. Attention or salience? Curr Opin Psychol. 2019;29:1-5.
43. Parr T, Rikhye RV, Halassa MM, Friston KJ. Prefrontal Computation as Active Inference. Cereb Cortex. 2020;30(2):682-95.
44. Chekroud AM. Unifying treatments for depression: an application of the Free Energy Principle. Front Psychol. 2015;6:153.
45. Kube T, Schwarting R, Rozenkrantz L, Glombiewski JA, Rief W. Distorted Cognitive Processes in Major Depression: A Predictive Processing Perspective. Biol Psychiatry. 2020;87(5):388-98.
46. Slavich G, Irwin M. From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. Psychological bulletin. 2014;140:774-815.
47. Smith R, Alkozei A, Killgore WDS, Lane RD. Nested positive feedback loops in the maintenance of major depression: An integration and extension of previous models. Brain Behav Immun. 2018;67:374-97.
48. Haker H, Schneebeli M, Stephan K. Can Bayesian Theories of Autism Spectrum Disorder Help Improve Clinical Practice? Frontiers in Psychiatry. 2016;7:107.
49. Lawson R, Rees G, Friston K. An aberrant precision account of autism. Frontiers in Human Neuroscience. 2014;8:302.
50. Van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, de-Wit L, et al. Precise minds in uncertain worlds: predictive coding in autism. Psychol Rev. 2014;121(4):649-75.
51. Corlett PR, Horga G, Fletcher PC, Alderson-Day B, Schmack K, Powers AR, 3rd. Hallucinations and Strong Priors. Trends Cogn Sci. 2019;23(2):114-27.
52. Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. Front Psychiatry. 2013;4:47.
53. Stephan K, Manjaly Z, Mathys C, Weber L, Paliwal S, Gard T, et al. Allostatic Self-efficacy: A Metacognitive Theory of Dyshomeostasis-Induced Fatigue and Depression. Frontiers in human neuroscience. 2016;10:550.
54. Smith R, Parr T, Friston KJ. Simulating Emotions: An Active Inference Model of Emotional State Inference and Emotion Concept Learning. Front Psychol. 2019;10:2844.
55. Lane RD, Weihs KL, Herring A, Hishaw A, Smith R. Affective agnosia: Expansion of the alexithymia construct and a new opportunity to integrate and extend Freud's legacy. Neurosci Biobehav Rev. 2015;55:594-611.
56. Smith R, Kaszniak AW, Katsanis J, Lane RD, Nielsen L. The importance of identifying underlying process abnormalities in alexithymia: Implications of the three-process model and a single case study illustration. Conscious Cogn. 2019;68:33-46.
57. Smith R, Killgore WDS, Lane RD. The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. Emotion. 2018;18(5):670-92.
58. Burger A, Lumley M, Carty J, Latsch D, Thakur E, Hyde-Nolan M, et al. The effects of a novel psychological attribution and emotional awareness and expression therapy for chronic musculoskeletal pain: A preliminary, uncontrolled trial. Journal of psychosomatic research. 2016;81:1-8.

59. Thakur E, Holmes H, Lockhart N, Carty J, Ziadni M, Doherty H, et al. Emotional awareness and expression training improves irritable bowel syndrome: A randomized controlled trial. Neurogastroenterology & Motility. 2017;29:e13143.
60. Ferry AL, Hespos SJ, Waxman SR. Categorization in 3- and 4-month-old infants: an advantage of words over tones. Child development. 2010;81:472-9.
61. Widen S, Russell J. Children acquire emotion categories gradually. Cognitive Development. 2008;23:291-312.
62. Gergely G, Watson J. The social biofeedback theory of parental affect-mirroring: the development of emotional self-awareness and self-control in infancy. The International Journal of Psychoanalysis. 1996;77:1181-212.
63. Smith R, Schwartenbeck P, Parr T, Friston KJ. An active inference approach to modeling concept learning. bioRxiv. 2019:633677.
64. Brown TH, Zhao Y, Leung V. Hebbian plasticity. Encyclopedia of Neuroscience2010. p. 1049-56.
65. Panksepp J, Lane RD, Solms M, Smith R. Reconciling cognitive and affective neuroscience perspectives on the brain basis of emotional experience. Neurosci Biobehav Rev. 2017;76(Pt B):187-215.
66. Smith R. The three-process model of implicit and explicit emotion. In: Lane R, Nadel L, editors. Neuroscience of Enduring Change: Implications for Psychotherapy: Oxford University Press; 2020.
67. Smith R, Alkozei A, Killgore WDS. How Do Emotions Work? Frontiers for Young Minds. 2017;5.
68. Smith R, Killgore WDS, Alkozei A, Lane RD. A neuro-cognitive process model of emotional intelligence. Biol Psychol. 2018;139:131-51.
69. Smith R, Lane RD. Unconscious emotion: A cognitive neuroscientific perspective. Neurosci Biobehav Rev. 2016;69:216-38.
70. Smith R, Lane RD. The neural basis of one's own conscious and unconscious emotional states. Neurosci Biobehav Rev. 2015;57:1-29.
71. Smith R, Alkozei A, Bao J, Smith C, Lane RD, Killgore WDS. Resting state functional connectivity correlates of emotional awareness. Neuroimage. 2017;159:99-106.
72. Smith R, Alkozei A, Lane RD, Killgore WDS. Unwanted reminders: The effects of emotional memory suppression on subsequent neuro-cognitive processing. Conscious Cogn. 2016;44:103-13.
73. Smith R, Bajaj S, Dailey NS, Alkozei A, Smith C, Sanova A, et al. Greater cortical thickness within the limbic visceromotor network predicts higher levels of trait emotional awareness. Conscious Cogn. 2018;57:54-61.
74. Smith R, Braden BB, Chen K, Ponce FA, Lane RD, Baxter LC. The neural basis of attaining conscious awareness of sad mood. Brain Imaging Behav. 2015;9(3):574-87.
75. Smith R, Fass H, Lane RD. Role of medial prefrontal cortex in representing one's own subjective emotional responses: a preliminary study. Conscious Cogn. 2014;29:117-30.
76. Smith R, Lane RD, Alkozei A, Bao J, Smith C, Sanova A, et al. The role of medial prefrontal cortex in the working memory maintenance of one's own emotional responses. Sci Rep. 2018;8(1):3460.
77. Smith R, Lane RD, Alkozei A, Bao J, Smith C, Sanova A, et al. Maintaining the feelings of others in working memory is associated with activation of the left anterior insula and left frontal-parietal control network. Soc Cogn Affect Neurosci. 2017;12(5):848-60.

78. Smith R, Lane RD, Sanova A, Alkozei A, Smith C, Killgore WDS. Common and Unique Neural Systems Underlying the Working Memory Maintenance of Emotional vs. Bodily Reactions to Affective Stimuli: The Moderating Role of Trait Emotional Awareness. Front Hum Neurosci. 2018;12:370.
79. Smith R, Sanova A, Alkozei A, Lane RD, Killgore WDS. Higher levels of trait emotional awareness are associated with more efficient global information integration throughout the brain: a graph-theoretic analysis of resting state functional connectivity. Soc Cogn Affect Neurosci. 2018;13(7):665-75.
80. Barrett L. How emotions are made: The secret life of the brain. 2017.
81. Mueller J, Alpers GW. Two facets of being bothered by bodily sensations: anxiety sensitivity and alexithymia in psychosomatic patients. Comprehensive Psychiatry. 2006;47:489-95.
82. Colvert E, Rutter M, Kreppner J, Beckett C, Castle J, Groothues C, et al. Do Theory of Mind and Executive Function Deficits Underlie the Adverse Outcomes Associated with Profound Early Deprivation?: Findings from the English and Romanian Adoptees Study. Journal of Abnormal Child Psychology. 2008;36:1057-68.
83. Fries AB, Pollak SD. Emotion understanding in postinstitutionalized Eastern European children. Dev Psychopathol. 2004;16(2):355-69.
84. Barrett L, Gross J, Christensen T, Benvenuto M. Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. Cognition & Emotion. 2001;15:713-24.
85. Kashdan T, Barrett L, McKnight P. Unpacking Emotion Differentiation: Transforming Unpleasant Experience by Perceiving Distinctions in Negativity. Current Directions in Psychological Science. 2015;24:10-6.
86. Lane RD, Anderson FS, Smith R. Biased Competition Favoring Physical Over Emotional Pain: A Possible Explanation for the Link Between Early Adversity and Chronic Pain. Psychosom Med. 2018;80(9):880-90.
87. Smith R, Steklis HD, Steklis NG, Weihs KL, Lane RD. The evolution and development of the uniquely human capacity for emotional awareness: A synthesis of comparative anatomical, cognitive, neurocomputational, and evolutionary psychological perspectives. Biol Psychol. 2020;154:107925.
88. Hayes S, Strosahl K, Wilson K. Acceptance and commitment therapy: An experiential approach to behaviour change. 2003.
89. Segal Z, Teasdale J, Williams J. Mindfulness-Based Cognitive Therapy: Theoretical Rationale and Empirical Status. In: Hayes S, Follette V, Linehan M, editors. Mindfulness and acceptance: Expanding the cognitive-behavioral tradition. New York: Guilford Press; 2004. p. 45-65.
90. Greenberg L. Emotion-Focused Therapy: Theory and Practice. 2010.
91. Barlow D, Allen L, Choate M. Toward a Unified Treatment for Emotional Disorders - Republished Article. Behavior therapy. 2016;47:838-53.
92. Smith R, Khalsa SS, Paulus MP. An Active Inference Approach to Dissecting Reasons for Nonadherence to Antidepressants. Biol Psychiatry Cogn Neurosci Neuroimaging. 2019.
93. Smith R, Moutoussis M, Bilek E. Simulating the computational mechanisms of cognitive and behavioral psychotherapeutic interventions: Insights from active inference. PsyArXiv. 2020.
94. Yamashita Y, Tani J. Spontaneous prediction error generation in schizophrenia. PLoS One. 2012;7(5):e37843.

95. Idei H, Murata S, Chen Y, Yamashita Y, Tani J, Ogata T. A Neurorobotics Simulation of Autistic Behavior Induced by Unusual Sensory Precision. Computational psychiatry (Cambridge, Mass). 2018;2:164-82.
96. Idei H, Murata S, Yamashita Y, Ogata T. Homogeneous intrinsic neuronal excitability induces overfitting to sensory noise: A robot model of neurodevelopmental disorder. PsyArXiv. 2020.
97. Friston KJ. Precision Psychiatry. Biological psychiatry Cognitive neuroscience and neuroimaging. 2017;2(8):640-3.
98. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, et al. Uncertainty in perception and the Hierarchical Gaussian Filter. Front Hum Neurosci. 2014;8:825.
99. Diaconescu AO, Mathys C, Weber LAE, Kasper L, Mauer J, Stephan KE. Hierarchical prediction errors in midbrain and septum during social learning. Soc Cogn Affect Neurosci. 2017;12(4):618-34.
100. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, den Ouden HE, et al. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron. 2013;80(2):519-30.
101. Deserno L, Boehme R, Mathys C, Katthagen T, Kaminski J, Stephan KE, et al. Volatility Estimates Increase Choice Switching and Relate to Prefrontal Activity in Schizophrenia. Biol Psychiatry Cogn Neurosci Neuroimaging. 2020;5(2):173-83.
102. Cole DM, Diaconescu AO, Pfeiffer UJ, Brodersen KH, Mathys CD, Julkowski D, et al. Atypical processing of uncertainty in individuals at risk for psychosis. Neuroimage Clin. 2020;26:102239.
103. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. Science. 2017;357(6351):596-600.
104. Adams R, Perrinet L, Friston K. Smooth Pursuit and Visual Occlusion: Active Inference and Oculomotor Control in Schizophrenia. PLoS ONE. 2012;7:e47502.
105. Lawson R, Mathys C, Rees G. Adults with autism overestimate the volatility of the sensory environment. Nature Neuroscience. 2017;20:1293-9.
106. Smith R, Schwartenbeck P, Stewart JL, Kuplicki R, Ekhtiari H, Investigators T, et al. Imprecise Action Selection in Substance Use Disorder: Evidence for Active Learning Impairments When Solving the Explore-exploit Dilemma. Drug and Alcohol Dependence. 2020:(in press).
107. Smith R, Kirlic N, Touthang J, Yeh HW, Kuplicki R, Khalsa SS, et al. Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: a computational modeling approach. Journal of Psychiatry & Neuroscience. 2020:(in press).
108. Smith R, Kuplicki R, Feinstein J, Forthman KL, Stewart JL, Paulus MP, et al. An active inference model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. medRxiv. 2020:2020.06.03.20121343.
109. Park HJ, Friston K. Structural and functional brain networks: from connections to cognition. Science. 2013;342(6158):1238411.
110. Bogacz R. A tutorial on the free-energy framework for modelling perception and learning. Journal of mathematical psychology. 2017;76:198-211.

Figure Legends

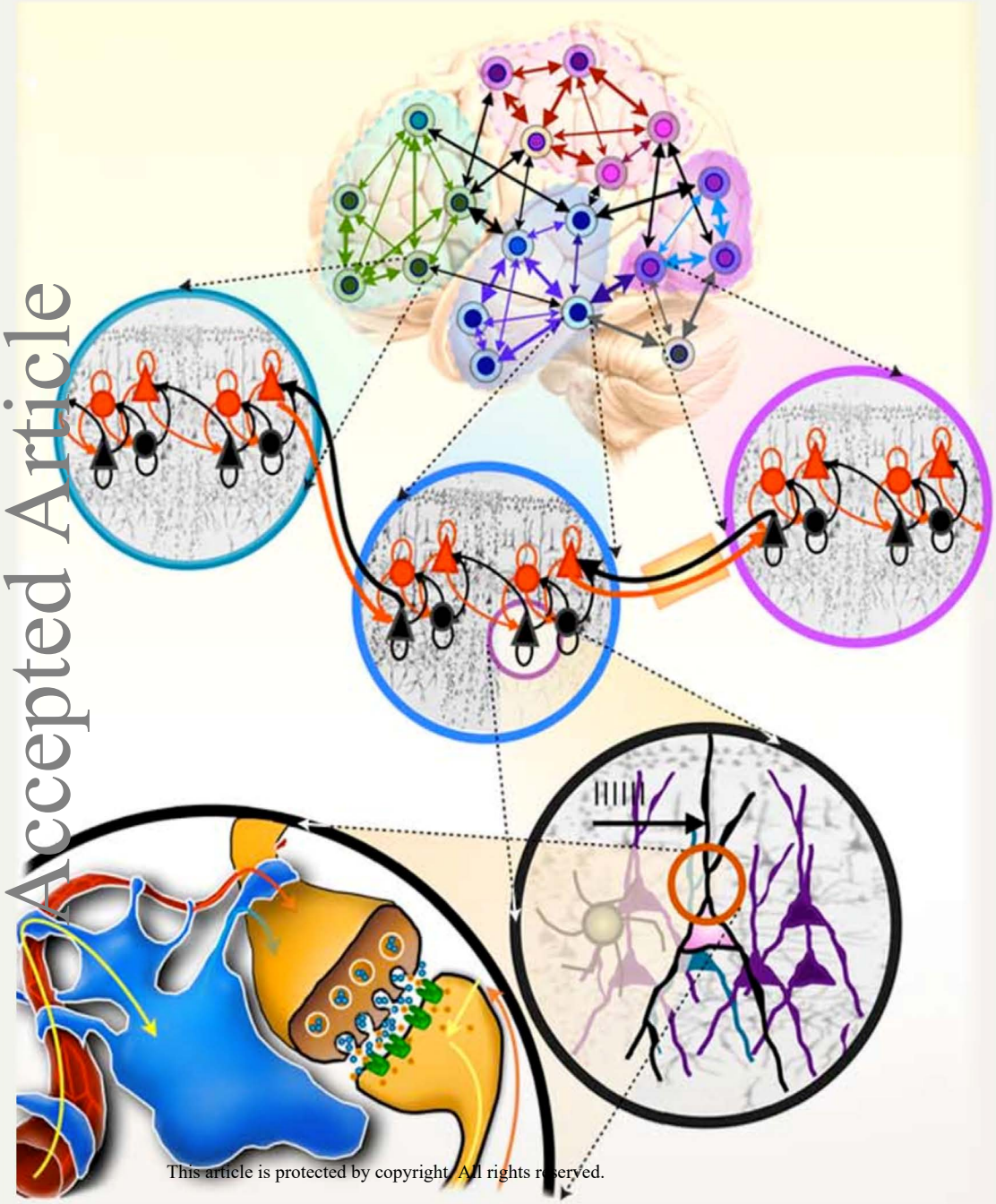
Figure 1. Schematic of the multi-scale architecture of neural networks. At the macro-scale, the brain is most often studied in terms of activity within, or interactions between, large-scale brain regions using neuroimaging. These brain regions can interact over large distances, mediated by long-range axonal fiber bundles. At the micro-scale, the brain is studied in terms of single-neuron activity and intra-/inter-cellular interactions at the molecular level. The meso-scale links the micro- and macro-scales, but it is currently the most difficult to study. Computational neuroscience is uniquely suited to address meso-scale function. Current work in this area has led to strong support for the idea that patterns of synaptic connectivity at this level implement the predictive algorithms discussed in the text. This algorithmic level of description bridges the neural and psychological levels of description, by showing how neural structures can implement the computational processes underlying psychological functions, such as perception, learning, and decision-making. Meso-scale computational (algorithmic) modelling therefore offers a unique window into the neural basis of abnormal psychological processes within psychopathology. Adapted from (109).

Figure 2. (**A**) An illustration of neuronal dynamics within a very simple predictive coding model of 'friendliness perception' (i.e., estimating how friendly vs. hostile someone is based on lip curvature, as an indicator of facial expression). This example illustrates how the same input (flat lip curvature, most consistent with a neutral facial expression) can be interpreted as a sign of higher or lower levels of friendliness based on different learned prior expectations and precision estimates. In this model, blue triangles indicate cortical pyramidal neurons, and black lines indicate axons terminating in synaptic connections. Arrows indicate excitatory synaptic influences, and circles indicate inhibitory synaptic influences (dashed arrows are not modeled, but indicate additional context-specific modulatory influences that could also be present in a more complete model). Activity of the FL neuron estimates level of friendliness (from very hostile to very friendly; activity levels from 0-10), and LC neuron activity represents lip curvature (i.e., low activity indicates frown and high activity indicates smile). The two PE neurons reflect prediction-errors associated with FL activation (higher level) and with LC activation (lower level). The strength of the two looping axons' synapses (connecting each PE neuron to itself) estimates the precision (reliability) of prior expectations (π_{FL}; higher level) and LC activity (π_{LC}; lower level). Expected friendliness (Pr_{FL}) is conveyed through the strength of the top-down inhibitory synapse on the higher-level PE neuron. Although not modeled here, predictive coding models also include quantitative synaptic learning mechanisms (i.e., update equations) allowing the strengths of the Pr_{FL}, π_{FL}, and π_{LC} synapses (i.e., prior expectations and precision estimates) to be altered over time to better match patterns in experience. Panels B-E illustrate changes in FL neuron activity (i.e. inferred friendliness level; black lines) over time, when presented with a flat lip curvature (i.e., moderate LC

activity, most consistent with a neutral level of friendliness, all else being equal; blue lines) under different model parameter values. These different parameter values reflect 1) prior expectations of low (**B**) vs. high (**C**) levels of friendliness, and 2) high (**D**) vs. low (**E**) reliability/precision estimates for expectations of low friendliness. As can be seen, after FL neuron activity converges onto a stable estimate (i.e., when activity ceases to oscillate once prediction error is minimized), lower levels of friendliness are inferred in (**B**) compared to (**C**) (reflecting the influence of prior predictions), and in (**D**) compared to (**E**) (reflecting the influence of higher reliability estimates for prior predictions of low friendliness). For the detailed mathematics on which this example is based, see (110). Panel (**F**) depicts the neural process theory proposed within active inference. Example simulations using this process theory are shown in **Figure 3**. In this theory, probability estimates for a given phenomenon (e.g., being friendly or hostile) are associated with neuronal populations that are arranged to reproduce known intrinsic (within cortical area) connections. Red connections are excitatory, blue connections are inhibitory, and green connections are modulatory (i.e., involve a multiplication or weighting). Similar to predictive coding, these connections convey different types of prediction and prediction error signals (labeled as signal types 1-5), but in this case also incorporating action selection. Cyan units correspond to predictions about future sensory inputs (o) and the causes of those inputs (s) if one were to follow one sequence of actions vs. another (i.e., policies, π) at each time point (t), while red units represent a type of 'best guess' about causes of sensory input when considering the probability of all possible policies (i.e., a Bayesian model average). Pink units correspond to sensory prediction errors (ε) and preference predictions errors (ζ), which are used to evaluate expected free energy (G) and subsequent policy probabilities (π). When selecting an action (u) at each time point, policy probabilities are also modulated by an expected precision term (γ) that has been linked to dopamine (20). Predictions are conveyed by a set of synaptic connections (B), while prediction-errors are encoded by a different set of synaptic connections (A); these synaptic connection strengths are updated during associative learning (i.e., long-term synaptic potentiation and depression; (64)). Only exemplar connections are shown to avoid visual clutter. Furthermore, we have just shown neuronal populations encoding beliefs under two possible policies over three time points. For an introduction to the associated mathematics describing neuronal interactions in this theory, see (1, 13, 15).

Figure 3. Depiction of a formal model of emotional state inference, adapted from (19, 54); also see (18). On the left, a network of abstract neuron-like nodes (and select connections) is depicted, where the higher-level represents different possible concept-level inferences about one's own internal state, including emotional and somatic interpretations. These representations send downward prediction signals, conveying patterns of lower-level representations expected under each higher-level representation. Prediction-errors are conveyed upward, such that the higher level converges onto the representation that makes the most accurate predictions (i.e., minimizes prediction error). The lower level in turn generates affective (interoceptive/behavioral) responses themselves (e.g., generating

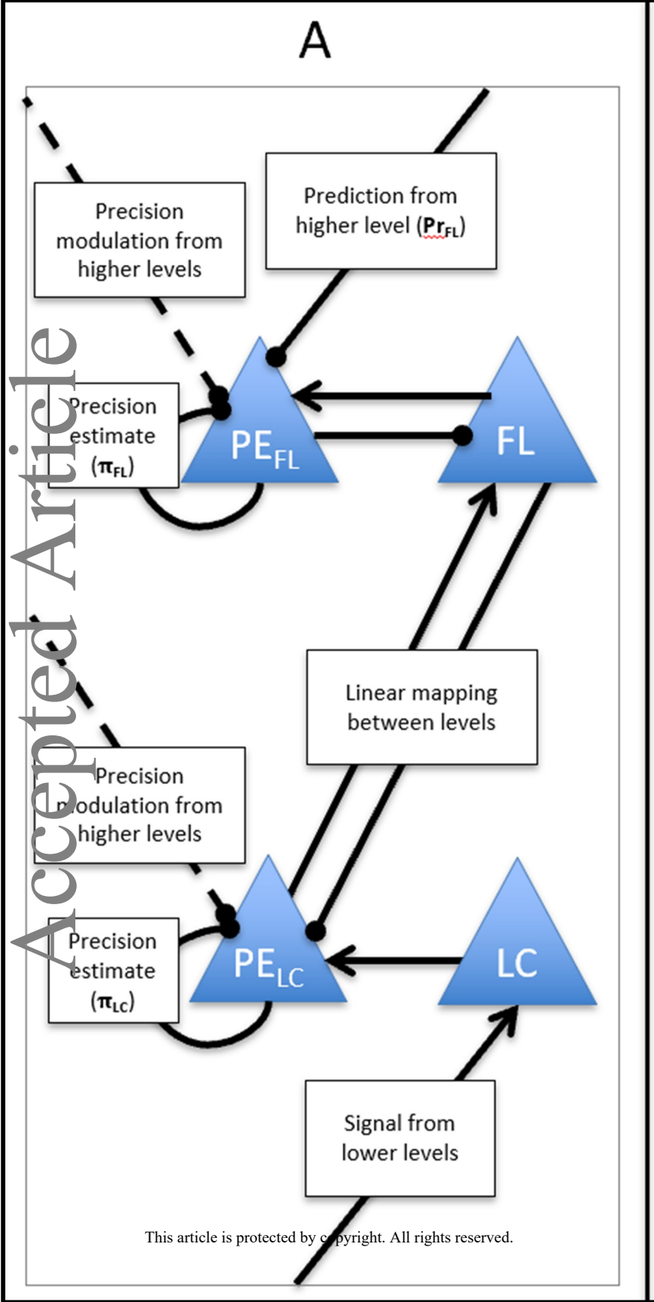
elevated heart rate and avoidance behavior when inferring threat; shown in gray). Here it is assumed that the brain has previously generated a (flexible, context-specific) affective response of this kind, based on the predicted metabolic, cognitive, and behavioral demands of a situation. This occurs upon perceiving and interpreting that situation (i.e., based on a fast, largely unconscious, domain-general evaluation of a situation as being, for example, safe, threatening, goal-incongruent, consistent with norms/values, etc.; only interpretations of the situation as neutral, socially threatening, or physically threatening are explicitly depicted here). The brain's job is then to perceive how body states have changed based on this generated response, and to make sense of it in terms of various (self-report guiding) concept representations it has learned (i.e., this is based on the three-process model of emotional awareness (57, 65-70), as well as constructivist theories stressing the culture- and context-specific, domain-general nature of emotion conceptualization (80)). To do so, the simulated decision-making 'agent' must first choose selective attention policies (i.e., selectively attend to valence, arousal level, the surrounding context, etc.) and then choose self-report policies communicating which emotion it is feeling, based on what it has attended to and observed (with the preference of receiving confirmatory social feedback, which also allows simulation of learning emotion concepts during development (54)). In this example, activated lower-level representations are shown in red, leading to the inference of sadness at the higher level (also red) and the self-report that 'I feel sad' (not shown). This occurs gradually as the individual selectively attends to each lower-level modality (not shown) and accumulates evidence for the sadness interpretation. After arriving at this interpretation, it can also be held in working memory and reflected upon (i.e., gestured at with the 'Domain-general Cognition' box in the top-left; see (19) for explicit simulations). The top-right corresponds to simulated neuronal firing rates (darker = higher firing rate; i.e., higher represented probability) and simulated local field potentials (rates of change in firing rates) for this model, based on the neural process theory proposed within the active inference framework (15), which is depicted in **Figure 2**. Here, the network starts with equally strong predictions across all possible interpretations, and slowly converges onto the belief that sadness is the best-fit interpretation when successively integrating lower-level representations (via selective attention processes not depicted here). By simulating different starting predictions and connection strengths, precise simulations of pathological emotion inference and learning can be carried out and can be used to make empirical predictions.

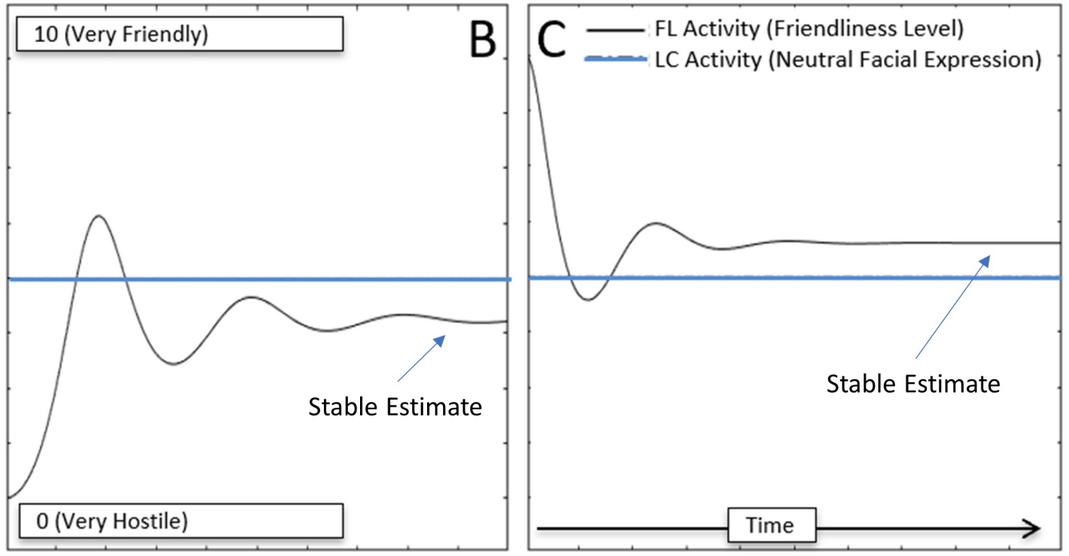


Macro-scale

Meso-scale

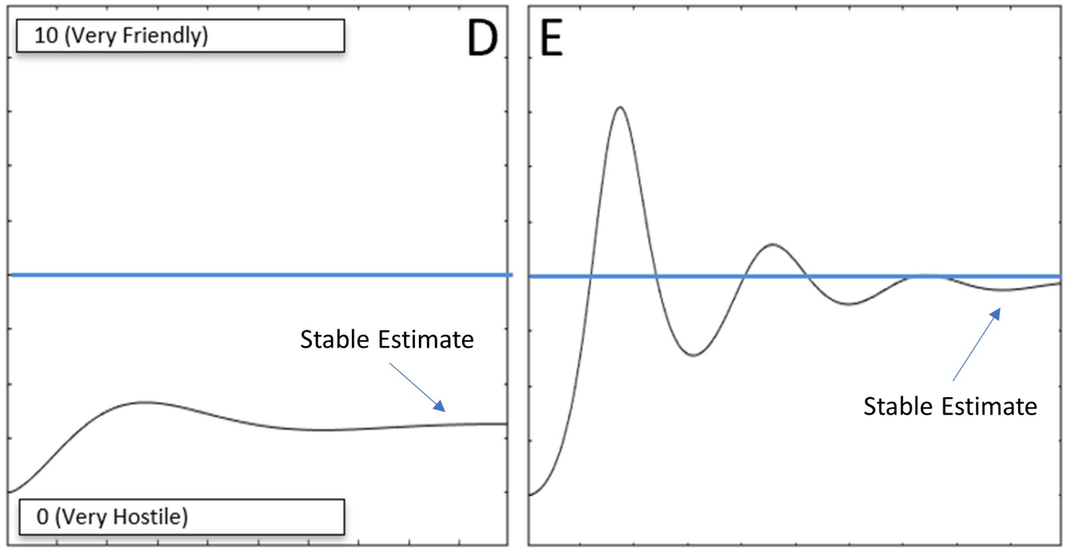
Micro-scale





Pr_{FL}= 1 (Low); $\pi_{FL} = \pi_{LC}$; LC= 5

Pr_{FL}= 9 (High); $\pi_{FL} = \pi_{LC}$; LC= 5



Pr_{FL}= 1 (Low); $\pi_{FL} > \pi_{LC}$; LC = 5

Pr_{FL}= 1 (Low); $\pi_{FL} < \pi_{LC;}$; LC = 5