Smith, R., Varshney, L. R., Nagayama, S., Kazama, M., Kitagawa, T., Managi, S., & Ishikawa, Y. (2022). A computational neuroscience perspective on subjective wellbeing within the active inference framework. *International Journal of Wellbeing*, *12*(4), 102-131. https://doi.org/10.5502/ijw.v12i4.2659

ARTICLE

A computational neuroscience perspective on subjective wellbeing within the active inference framework

Ryan Smith · Lav R. Varshney · Susumu Nagayama · Masahiro Kazama
Takuya Kitagawa · Shunsuke Managi · Yoshiki Ishikawa

Abstract: Understanding and promoting subjective wellbeing (SWB) has been the topic of increasing research, due in part to its potential contributions to health and productivity. To date, the conceptualization of SWB has been grounded within social psychology and largely focused on self-report measures. In this paper, we explore the potentially complementary tools and theoretical perspectives offered by computational neuroscience, with a focus on the active inference (AI) framework. This framework is motivated by the fact that the brain does not have direct access to the world; to select actions, it must instead infer the most likely external causes of the sensory input it receives from both the body and the external world. Because sensory input is always consistent with multiple interpretations, the brain's internal model must use background knowledge, in the form of prior expectations, to make a "best guess" about the situation it is in and how it will change by taking one action or another. This best guess arises by minimizing an error signal representing the deviation between predicted and observed sensations given a chosen action—quantified mathematically by a variable called free energy (FE). Crucially, recent proposals have illustrated how emotional experience may emerge within AI as a natural consequence of the brain keeping track of the success of its model in selecting actions to minimize FE. In this paper, we draw on the concepts and mathematics in AI to highlight how different computational strategies can be used to minimize FE—some more successfully than others. This affords a characterization of how diverse individuals may adopt unique strategies for achieving high SWB. It also highlights novel ways in which SWB could be effectively improved. These considerations lead us to propose a novel computational framework for understanding SWB. We highlight several parameters in these models that could explain individual and cultural differences in SWB, and how they might inspire novel interventions. We conclude by proposing a line of future empirical research based on computational modelling that could complement current approaches to the study of wellbeing and its improvement.

Keywords: subjective wellbeing, active inference, computational neuroscience, computational psychiatry, predictive coding, emotion

1. Introduction

Understanding and improving subjective wellbeing (SWB) is important in promoting resilience, social support, physical health and longevity, work performance, and broader contributions to society (De Neve et al., 2013; Diener et al., 2017). Previous research has shown that individuals reporting higher SWB tend to be healthier and live longer, due in part to more frequent

Ryan Smith
Laureate Institute for Brain Research
rsmith@laureateinstitute.org

engagement in health behaviors (Danner et al., 2001; Diener & Chan, 2011; Lyubomirsky et al., 2005; Steptoe & Wardle, 2011). They also have more successful marriages (Lucas et al., 2003; Luhmann et al., 2013), more friendships (Moore et al., 2018), and both greater reported job satisfaction and better job performance (Borman et al., 2001; Tenney et al., 2016). In addition, they return to healthy emotional states more quickly after negative life events (Fredrickson et al., 2003). Although much of the research in this area is correlational, there are also established mechanisms through which emotional health can promote physical health (Slavich & Irwin, 2014), and arguments can be made that causal influences between each of the aforementioned measures are synergistic and bidirectional (Diener et al., 2018).

To date, the primary measures used to study SWB—and which have led to the important findings discussed above—are based on self-report (Diener et al., 1985; Diener et al., 2018; Diener et al., 2010; Pavot et al., 1991; Sandvik et al., 1993; Weziak-Bialowolska et al., 2021). In a few studies, this approach has been supplemented with other methods, such as third-party reports from friends, family, and co-workers, with results supporting convergent validity (Pavot et al., 1991; Sandvik et al., 1993; Schimmack & Oishi, 2005; Schneider & Schimmack, 2009); for other work supporting combinations of self-report and quantitative measures (e.g., the Human Development Index), see (Anand & Sen, 1992; Sen, 1985). In a small number of studies, report-based SWB measures have also been associated with objective measures such as smiling intensity (Seder & Oishi, 2012), the emotional valence of word choice (Schwartz et al., 2013), and peripheral physiological measures of stress and health (Steptoe et al., 2005). There have also been efforts to improve SWB through interventions focused on exercise, mindfulness, and/or metacognitive training (reviewed in (Varshney & Barbey, 2021)). However, conventional descriptive approaches within SWB research have been limited in their ability to provide insights beyond associations between variables of interest. As such, the mechanistic and biobehavioral basis of SWB remains poorly characterized and our understanding could be greatly advanced by applying complementary approaches in neighboring fields.

One promising set of complementary approaches comes from computational neuroscience, and its clinically-focused sister discipline of computational psychiatry. Various lines of work in these fields have focused on understanding the mechanistic basis of individual differences in both pathological and sub-clinical levels of emotional symptoms—such as depression/anxiety and the way they interfere with overall life functioning—with clear connections to differences in SWB (Friston et al., 2014; Huys et al., 2016; Montague et al., 2012). Computational approaches focus on mathematical models of brain processes to explain complex patterns of learning/behavior and may therefore offer a complementary approach for improving understanding of the neurocognitive mechanisms underlying previously observed differences in SWB based on self-report.

As we will describe further below, these approaches may offer at least two specific advantages to the field of SWB research. One is that they facilitate construction of general theories with broad explanatory power. For example, constructivist and descriptive perspectives have highlighted how conceptualizations of emotion and wellbeing can differ based on cultural and historical context (Barrett, 2017; Lomas et al., 2021). However, once such differences are identified, conventional descriptive approaches in SWB research may have difficulty explaining why these differences exist. In contrast, if knowledge of the more general neurocomputational processes underlying cognition and behavior can be incorporated, the origin of such differences may be accounted for within a single model (e.g., based on specific learning mechanisms or inference processes). A second advantage is that mathematical models of cognition offer an additional type of predictive ability. This is because they allow for quantitative simulation of non-trivial counterfactual changes in cognition/behavior under different circumstances. For example, one could simulate (and therefore predict) how someone would act if they were to have one set of experiences vs. another or if they were to have one set of beliefs vs. another. The predictions that emerge from such simulations can then be evaluated empirically. In turn, experimental results can refine available models before further rounds of simulation, prediction, and testing—forming in an iterative process of data-driven improvement of current theories. Importantly, this simulation approach can also be applied to hypothetical interventions. Namely, mathematical models can be used to simulate the outcomes of a proposed intervention and then studies can test whether the predicted outcomes occur.

Motivated by these possible advantages, in this paper we will first provide a brief introduction to computational approaches for wellbeing researchers and social psychologists who do not have background in this area. We describe the mathematics at a conceptual level and provide thorough explanation of any equations shown. We then discuss how computational approaches might yield novel lines of research on SWB. While we will touch upon multiple computational models, we focus primarily on the active inference framework and the additional resources it may offer, including: 1) novel theoretical conceptualizations of SWB, and 2) additional analytic tools that could be used to advance empirical research on the mechanisms and determinants of SWB (and behaviors that promote SWB). As we describe in more detail below, this approach affords behavioral measures of individual differences in a range of neurocomputational processes, including those underlying flexibility in the future-directedness of planning, motivation to seek out information, beliefs about environmental predictability, the influence of expectation on perception, and the sophistication with which individuals understand emotions, among others. Gathering such information will allow assessment of the mechanisms contributing to differences in SWB and could inspire novel interventions for improving SWB by targeting those mechanisms.

2. Computational approaches and conceptual links to wellbeing

2.1 Reinforcement learning

One widely established branch of computational neuroscience is reinforcement learning (RL). Within RL, it is assumed that individuals make decisions to maximize expected reward. One way they can do so is by maintaining an internal model of the world (so-called "model-based" RL), and using that model to plan the sequence of actions that would maximize cumulative reward (this type of internal model can also be used to facilitate learning through internal simulation; (Sutton & Barto, 1998)). Because model-based RL is computationally expensive (and in some cases intractable), many applications have instead focused on "model-free" RL, where individuals learn by trial-and-error without making explicit predictions about the future situations that will arise[1]. Instead, they simply learn an expected reward value for each possible action in a given situation and then choose the action with the highest expected value (often with a certain amount of randomness built in). Learning in these models is based on so-called "reward prediction errors" (RPEs). Briefly, when the expected reward following an action does not match observed reward, this mismatch (prediction error) updates the expected reward value of that action (increasing it if better than expected, decreasing it if worse than expected). We can calculate this prediction error as follows:

$$RPE = Observed\ Reward - Expected\ Reward$$

The expected reward of the action that led to the RPE is then updated based on the following rule:

$$New\ Expected\ Reward = Old\ Expected\ Reward + (RPE \times Learning\ Rate)$$

Here the learning rate is a value that controls how strong the change in expected reward is after each RPE. In tasks expected to have stable reward probabilities, learning rates should be low (since unexpected rewards may just be low-probability events). In contrast, if reward probabilities are expected to change every so often (i.e., the environment is "volatile"), learning rate should be high, since unexpected rewards are more likely to indicate that the underlying reward probabilities have changed. If one can quantify an individual's learning rate, this may therefore provide information about how stable/predictable they expect the world to be (e.g., a low learning rate would entail the belief that previously learned action-outcome probabilities are unlikely to change).

To date, some work within the RL framework has linked emotional states and moods with patterns of recent rewards and RPEs (Blain & Rutledge, 2020; Eldar & Niv, 2015; Eldar et al., 2018; Eldar et al., 2016; Mason et al., 2017; Rutledge et al., 2014, 2015; Vanhasbroeck et al., 2021). For example, greater momentary happiness during reward learning tasks has been associated with stronger positive RPEs in recent trials (Rutledge et al., 2014; Vanhasbroeck et al., 2021). This relation to momentary happiness also appears to be related to how surprising a reward is, rather than the magnitude of reward per se; (Blain & Rutledge, 2020)). That is, individuals report greater momentary increases in happiness if they are more surprised that a reward was received. Individuals also report greater momentary happiness in probabilistic learning tasks with more stable/predictable reward probabilities (i.e., those with lower volatility), whereas tasks in which probabilities change unpredictably are instead associated with greater self-reported negative emotion and stress (Blain & Rutledge, 2020; de Berker et al., 2016). This line of work has therefore provided important insights. However, these studies have focused primarily on momentary happiness, as opposed to the computational basis of general life satisfaction or overall life functioning. Thus, more research is needed to understand the relationship between these computational reward learning processes and established SWB measures.

[1] There are also several hybrid or intermediate algorithms that have been proposed, such as two-system architectures (where model-based and model-free systems cooperate/compete) and successor representation algorithms, among others (e.g., appealing to the role of long-term memory).

2.2 Active inference

Another conceptually rich computational framework, which we focus on here, is *active inference* (AI; see **Figure 1** below). This framework describes how individuals select sequences of actions (called "policies", denoted by the variable π) based on their expected sensory consequences (often called observations or outcomes, denoted by the variable o). These observations can include rewards as well as other sensory inputs capable of reducing uncertainty about the underlying states of the world that must be inferred (denoted by the variable s) (Smith, Friston, et al., 2022). For example, hearing a barking sound (observation) may allow one to infer that a dog is most likely nearby (state), although one cannot see the dog directly. Unlike model-free RL, this type of decision process requires the brain to maintain an internal model of the world that can simulate predicted patterns of observations under different possible actions. It can then select the actions expected to generate the most preferred/informative outcomes. Crucially, behavior will quickly become maladaptive if predictions become inaccurate. Therefore, the brain's internal model must also be continually updated and revised in light of new sensory input. The AI framework suggests that this is accomplished by finding an updated set of beliefs after each observation that minimizes prediction error. Unlike RL, however, this is not an RPE. Instead, this more general prediction error can reflect a deviation between any predicted and observed sensory input. Importantly, however, there are typically multiple sets of new beliefs that could minimize prediction error, which means some further criterion is also needed to select among those possibilities. This criterion corresponds to parsimony; namely, beliefs should minimize prediction error while also *changing as little as possible*. Put another way, the brain should identify the simplest possible (or least "complex") change in belief that is necessary.

To capture this, AI formally models the brain as seeking to minimize a quantity called variational free energy (VFE), which represents a trade-off between minimizing prediction error and minimizing the complexity of change in belief. Without introducing the detailed mathematics, we can represent this as:

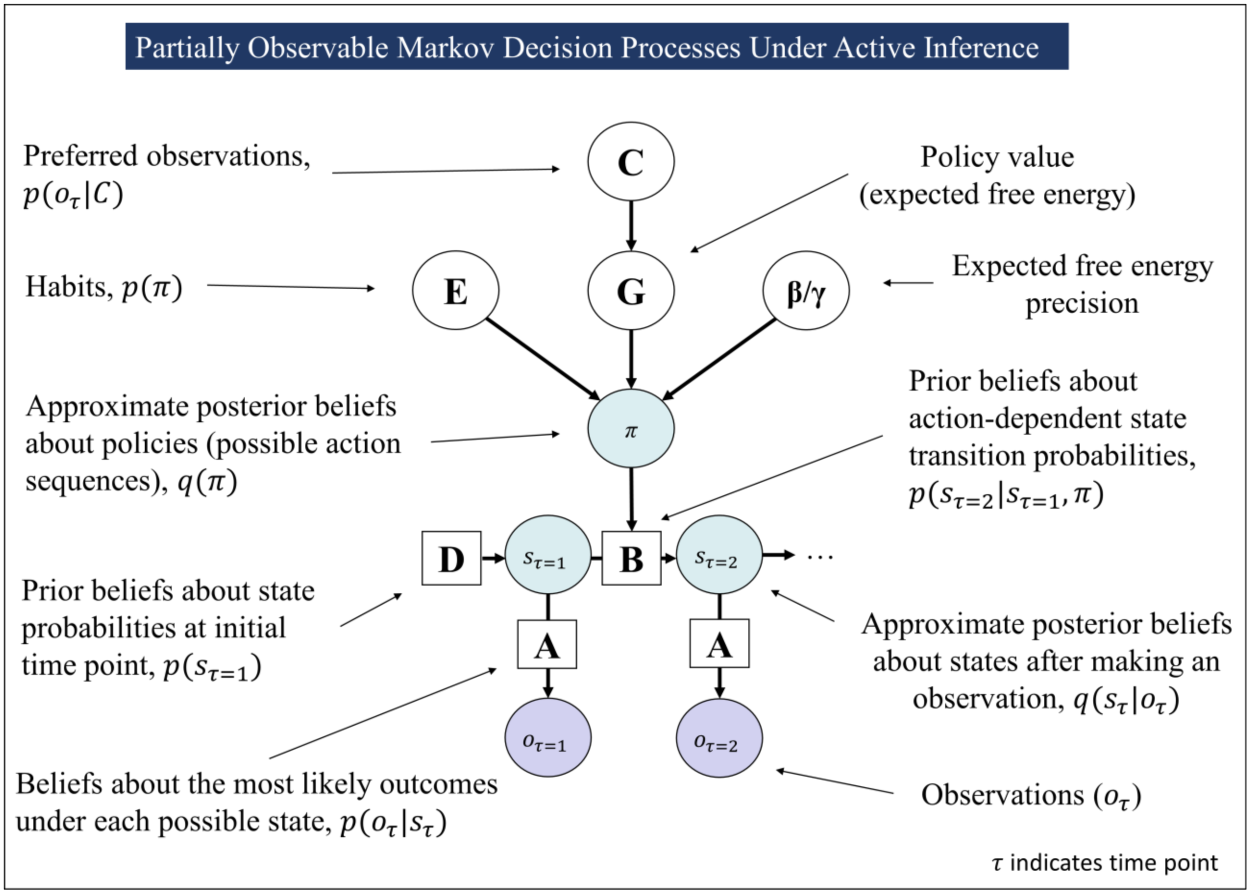
$$VFE = Complexity + Prediction\ Error$$

Importantly, previous literature has suggested that negative emotion may be a consequence of a chronic failure to successfully minimize VFE (or simply prediction error; e.g., see (Barrett et al., 2016; Joffily & Coricelli, 2013; Stephan et al., 2016)). This suggests that successful minimization of this quantity could also correspond to high SWB. However, this has not been empirically tested (although we discuss studies that have tested related hypotheses below).

While VFE relates beliefs to current observations, selecting actions requires the prediction of future observations ("what will I observe if I do this or that?"). This means the brain must evaluate *expected* free energy (EFE; denoted by variable **G** in **Figure 1**). Conceptually, minimizing EFE can be described as selecting actions expected to maximize both reward and information gain. This can be represented as:

$$EFE = -Epistemic\ Value - Expected\ Reward$$

Figure 1. Graphical depiction of active inference models.


Partially Observable Markov Decision Processes Under Active Inference
Preferred observations, $p(o_\tau|C)$
C
Policy value (expected free energy)
Habits, $p(\pi)$
E
G
β/γ
Expected free energy precision
Approximate posterior beliefs about policies (possible action sequences), $q(\pi)$
π
Prior beliefs about action-dependent state transition probabilities, $p(s_{\tau=2}|s_{\tau=1},\pi)$
D
$s_{\tau=1}$
B
$s_{\tau=2}$
…
Prior beliefs about state probabilities at initial time point, $p(s_{\tau=1})$
A
A
Approximate posterior beliefs about states after making an observation, $q(s_\tau|o_\tau)$
$o_{\tau=1}$
$o_{\tau=2}$
Beliefs about the most likely outcomes under each possible state, $p(o_\tau|s_\tau)$
Observations (o_τ)
τ indicates time point

Notes. Arrows denote asymmetric dependencies (e.g., the expected observation, o, depends on the underlying state of the world, s). The general notation $p(x)$ indicates the known probability for each possible value of some variable x, whereas $q(x)$ denotes an approximation or "best guess" about those probabilities (also referred to as a "posterior" belief, because this guess is updated after one makes a new observation). The notation $p(x|y)$ denotes the probability of each possible value of some variable x after one has learned the value of some other variable y. See text for further explanation. As illustrated here, the underlying states of the world change over time, and the way they change can depend on one's chosen sequence of actions (π). Actions are chosen based on a combination of habits, the expected free energy, and beliefs about the reliability or "precision" of expected free energy (described further in the text). After each new observation, an individual can use this model to come up with a best guess about how the state of the world has changed (based on minimizing variational free energy or prediction error) and then select actions expected to maximize both reward and information gain (based on minimizing expected free energy). For a detailed walkthrough of the mathematics, see (Smith, Friston, et al., 2022). Minimizing both types of free energy can be conceptualized as maximizing SWB, and the success with which one does so will depend on the values of various parameters in the model. This is described further in the main text.

Here, the "epistemic value" term refers to how much one's uncertainty would be reduced by an expected observation (i.e., how much it would increase one's confidence in the underlying state of the world). The consequence of this equation is that individuals who do not know how to reach their goals will first explore their environment to gain more information. Once uncertainty is resolved, actions will then become reward-seeking (i.e., seeking the observations with the highest values encoded in the preference parameters within **C** in **Figure 1**). Active inference models can also include other elements, such as learning rates, habits, and volatility beliefs, among others (for a summary of relationships between AI and RL, see **Table 1** below). However, their main distinguishing features from traditional RL approaches are: 1) prediction error (VFE) minimization as the basis of perception (traditional RL models do not include perception), and 2) information-seeking during action selection (although some more recent RL models have incorporated additional elements to drive strategic exploration; e.g., see (Gershman, 2018)). For a formal graphical depiction of active inference and some additional technical details, see Figure 1 and the associated legend. For a detailed walkthrough of the mathematics, see (Smith, Friston, et al., 2022).

Aside from the model structure depicted in **Figure 1**, there are also related model architectures based on free energy minimization (i.e., minimization of prediction errors weighted by their estimated reliability), with a primary example being the Hierarchical Gaussian Filter (HGF; (Mathys et al., 2014)). One advantage of the HGF is that it explicitly models volatility estimation (which then continuously updates learning rates after each new observation). In contrast, the simplest version of the AI architecture depicted in **Figure 1** only includes static learning rates. However, there are extensions to standard AI architectures that can perform explicit volatility estimation as well (Sales et al., 2019).

3. Potential relationships between free energy minimization and subjective wellbeing

There are fairly straightforward theoretical connections between EFE and SWB. If an individual continually fails to attain preferred observations (i.e., if EFE estimates continually fail to promote actions that achieve those observations), this would suggest they have a poor model of the world and are unable to improve it—perhaps resulting in feelings of helplessness and lack of control associated with low SWB. Previous theoretical work has also shown how the brain may keep track of the reliability/precision of its own EFE estimates (i.e., the parameter $\boldsymbol{\gamma}$ in **Figure 1**). In this case, a belief that EFE estimates are unreliable would be expected to promote negative affect (Hesp et al., 2021; Hesp et al., 2020)—because it entails an unsuccessful model of the world (i.e., a model that is not reliable in its ability to achieve preferred outcomes). Conversely, a consistent belief that one's model is successful at guiding action toward achieving one's goals could contribute to higher SWB. Two empirical studies to date have observed the expected correlation between negative affect (self-reported anxiety/uncertainty) and EFE reliability/precision estimates (using a model of participant behavior during an approach-avoidance conflict task (Smith, Kirlic, Stewart, Touthang, Kuplicki, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, McDermott, et al., 2021)), and three other studies using active inference models have linked related measures of precision to depression, anxiety, and/or substance use disorder severity (Smith, Kuplicki, Feinstein, et al., 2020; Smith, Schwartenbeck, et al., 2020; Smith, Taylor, Stewart, et al., 2022). One study has also identified neural correlates of EFE estimates using functional neuroimaging (Schwartenbeck et al., 2015). However, possible links to measures of SWB have not been examined.

Table 1. Computational frameworks that could be used to study subjective wellbeing.

	Reinforcement learning	**Active inference**
Explanatory targets	Cognitive and neural mechanisms of reward learning and reward-based decision-making	Cognitive and neural mechanisms of perception, learning, and decision-making
Guiding principle	Reward Maximization *Learning*: Updating beliefs about reward probabilities *Decision-making*: Maximizing cumulative rewards	Free energy principle Inferring beliefs about states of the world that minimize variational free energy (i.e., minimize complexity + prediction error) *Learning*: Inferring model parameters (e.g., probabilities of observations under different states) that improve the accuracy of model predictions (also based on variational free energy minimization). *Decision-Making*: Making choices that minimize expected free energy (i.e., which maximizes information gain + reward)
Example parameters	*Basic models*: State-space structure (granularity), learning rate, reward probabilities, reward sensitivity (level of randomness in choice) *Extended models*: Planning horizon, information bonus terms (driving information seeking)	Precision of: prior beliefs over states, state-observation mappings (sensory precision), preferences, expected free energy, and habits State-space structure (granularity), probabilities of observations given states, learning and forgetting rates; beliefs about environmental volatility; planning horizon
Hypotheses about the potential basis of subjective wellbeing	Greater momentary subjective wellbeing corresponds to patterns of repeatedly unexpected reward (i.e., better-than-expected outcomes)	Successful minimization of variational and expected free energy; high confidence in one's internal model of the world and its ability to adaptively guide action A match between one's internal model (i.e., model structure and parameter values) and the true statistics of the environment

Crucially, difficulties minimizing EFE can occur with respect to both internal bodily states (interoception) as well as external circumstances. For example, Stephan et al. (2016) have proposed that depressive symptoms develop due to a chronic failure to minimize interoceptive prediction error—associated with reduced confidence in the ability to successfully regulate the body and maintain preferred interoceptive observations (reduced "allostatic self-efficacy"). This also relates to interesting recent modelling work aiming to capture the computational basis of stress habituation—a phenomenon in which some individuals show reduced physiological responses to repeated presentation of the same stressor (Hartwig et al., 2022). This modelling work proposes that stress habituation can be captured as a reduction in the strength of the expected reward component of EFE (technically, a reduced precision of the distribution encoding preferred outcomes). In brief, when no course of action can be found to reduce EFE, a last resort option is to reduce the reward-seeking motivation itself. This means that the individual ceases to experience aversive stress responses (i.e., EFE is reduced), but at the cost of no longer expecting to achieve outcomes consistent with high SWB (e.g., remaining in unhealthy relationships or low socio-economic status environments, etc.). This computational model also introduces a parameter to account for the fact that some individuals show stress habituation and others do not. This highlights how there are trade-offs and individual differences in the way people may go about reducing EFE, not all of which are equally likely to promote high SWB.

While the theoretical relationship between SWB and EFE is fairly straightforward, possible links to VFE are more complex. Intuitively, one might simply equate higher SWB with lower levels of VFE (Joffily & Coricelli, 2013). At a biological level, this idea can be motivated by the fact that VFE minimization provides a generic means of describing successful survival and environmental adaptation (Friston, 2019; Kirchhoff et al., 2018). The mathematics behind this idea—referred to as the "free energy principle"—are beyond the scope of the present paper, but the key premise is that observations consistent with an organism's survival must be those expected to have the highest probability in that organism's model (e.g., if healthy levels of hydration were not observed with high probability, an organism would not survive). It follows that there will be large prediction errors (i.e., greater VFE) when survival-consistent outcomes are not observed. This formulation therefore entails that successful organisms act to minimize VFE, where this success at maintaining survival-consistent outcomes would be expected to promote higher SWB.

However, this is likely oversimplified when considering the study of human SWB. One basic reason for this is that minimizing VFE is also accomplished by arriving at new beliefs that minimize prediction error (in the simplest way possible)—and sometimes the most accurate beliefs can promote low SWB (e.g., believing that one is worthless may best minimize prediction error in some circumstances). Another reason is that even very simple organisms can be described as acting to minimize VFE as a means of survival, while there is substantially greater subjectivity and complexity associated with self-reported SWB in humans. This is because self-reported SWB depends on abstract beliefs about oneself, which requires an internal model of the world that can support this type of advanced cognition (Hesp et al., 2021; Hesp et al., 2020). Crucially, this can include abstract expectations about the conceptual requirements of SWB that go beyond, or even work against, biological fitness (e.g., some individuals may believe they need to be rich to be happy, while others do not; or some individuals may value self-sacrifice over personal safety, while others do not). Another complexity is that there are different concepts of wellbeing (e.g., hedonic vs. eudaimonic; (Ryan & Deci, 2001)). While momentary reductions in VFE might plausibly influence short-term hedonic wellbeing in some circumstances (e.g., when a return to homeostasis leads to positive emotion), eudaimonic wellbeing involves learned normative beliefs about what the requirements are for living (or having lived) a meaningful life that can counteract biological fitness (e.g., preferring states of "suffering in support of a meaningful cause" over states of "meaningless pleasure"). Given learned preferences for outcomes consistent with eudaimonic wellbeing, individuals will plausibly minimize EFE to realize them, but VFE could be minimized by inferring either high or low levels of eudaimonic wellbeing (i.e., depending on which inference best minimizes prediction error with respect to current observations). The role of inference through VFE minimization—whether those inferences promote eudaimonic wellbeing or not—is also exemplified by related work that has characterized determinants of feeling meaning in life (reviewed in (Kim et al., 2022)). This work suggests that several types of beliefs contribute to feelings of meaning, including: 1) that the various aspects of one's life are consistent and cohesive with one another; 2) that one's life has a purpose (it is working toward valued goals); and 3) that one's existence matters in the sense of having value, importance, and significance in the world. Meaning in life is also greater in those who show an intrinsic appreciation for various experiences (e.g., music, nature, interactions with family/friends; (Kim et al., 2022)). In active inference, the presence or absence of each of these beliefs would in part reflect inferences based on current observations—that is, the inferences that minimize VFE. However, some of these beliefs clearly also depend on expected future states and observations, which means EFE would also play a role similar to that described above. Levels of eudaimonic wellbeing will therefore reflect a complex mixture of VFE and EFE (with respect to beliefs about the present and future, respectively).

These considerations illustrate why the potential relationship between VFE minimization and SWB is not as straightforward as is the case with EFE. While VFE can be understood to track basic biological fitness, this is neither necessary nor sufficient for an individual to report high SWB. Rather, both hedonic and eudaimonic SWB are learned concepts used to evaluate one's experience, and self-reported SWB depends on how these learned concepts are encoded within the structure of an individual's internal model (e.g., an individual could potentially learn to associate low levels of hedonic SWB with high levels of eudaimonic SWB). The theoretical complexity here can also be highlighted by noting again how a model with beliefs about SWB will require hierarchical structure, which can allow VFE to be high at some levels and low at others. For example, consider an individual who values the idea of "hard work" and comes home feeling exhausted after a long day at their job. This feeling of exhaustion may reflect high VFE at a biological level (i.e., low metabolic resources); yet, that same feeling may lead to an elevated sense of self-worth (i.e., being exhausted provides evidence that one is a hard worker). This suggests that VFE minimization at some levels of processing may be more plausible correlates of SWB than others. However, despite these theoretical complexities, we will see below that—in the context of practical applications of active inference to SWB research—there may be several promising research directions to pursue.

4. Computational phenotypes: Multiple strategies for minimizing free energy

Based on the theoretical foundations discussed above, this section expands on the active inference framework and provides concrete examples of how SWB research could be advanced by applying computational concepts and approaches. While we focus on active inference due to the richer set of constructs it allows us to draw from, we note that—because minimizing EFE includes a reward maximization component with some mathematical similarities to RL (Da Costa et al., 2020)—some points made below may also apply to elements of that framework. One main message of this section is that there are likely multiple computational strategies that can successfully maintain high SWB. These strategies correspond to different internal model structures or "computational phenotypes" (Schwartenbeck & Friston, 2016), which can vary in the degree to which they adaptively capture the structure of the local environment (especially the local sociocultural environment). Computational phenotypes are defined in terms of the values of model parameters that best describe an individual's behavior. Below we will describe six different example model parameters, how they have been used in existing studies employing active inference models, and how they may explain differences in SWB.

4.1 Precision of prior beliefs over states

In a given context, an internal model—technically referred to as a "generative model" (because it generates predictions)—will include prior beliefs about what will be perceived (both dependent and independent of one's actions; encoded in parameters within **D** and **B** in **Figure 1**). If an individual is highly confident in their prior beliefs (if those beliefs are very "precise"), they will lead to a strong interpretive bias favoring expectations. Depending on the content domain, such prior beliefs could either promote or fail to promote SWB. For example, individuals with a precise prior belief that others tend to be friendly and supportive in social contexts may be more likely to interpret social signals in a positive manner—likely leading to actions that garner additional social support and promote greater SWB. Pessimistic prior biases, as in emotional disorders, can instead bias perception toward interpretations of threat and social rejection—generating chronic negative affect (Smith, Alkozei, et al., 2018). However, it is important to consider that, while an optimistic social interpretation bias could be helpful to a certain degree, it is also crucial that these expectations do not deviate too far from the true statistics of the environment. For example, a prior expectation that individuals are friendly could promote poor and unsafe choices in socially hostile environments and thus generate circumstances hindering SWB. Thus, there is not one optimal prior belief. It depends on the content domain and on a match with the environment. Model accuracy can be optimized so long as individuals also believe the sensory signals from the environment are reliable. If sensory signals are believed to be unreliable—referred to as having low "sensory precision" estimates (encoded within **A** in **Figure 1**)—prior beliefs will more strongly dominate perception and potentially promote false beliefs.

The combination of socio-behavioral tasks and computational modeling can measure the precision of prior beliefs in perception at the individual level, providing one dimension of each participant's computational phenotype. For examples of studies measuring individuals' prior perceptual beliefs and sensory precision estimates in other content domains, see (Powers et al., 2017; Smith, Kuplicki, Feinstein, et al., 2020; Smith, Kuplicki, Teed, et al., 2020; Smith, Mayeli, et al., 2021). To our knowledge, no study has measured prior beliefs or sensory precision estimates in relation to measures of SWB in social decision tasks (for related theoretical work linking imprecise prior beliefs to racial discrimination, see (Varshney & Varshney, 2016)). This therefore represents one promising avenue for future research.

4.2 Learning rates, forgetting, and volatility estimates

A second important dimension of a computational phenotype describing the way individuals try to minimize VFE/EFE corresponds to learning rate. Learning rates were briefly introduced above in the context of RL models, where greater learning rates imply a belief that previously learned action-outcome probabilities are unstable over time (i.e., the environment is highly volatile). If volatility estimation processes function appropriately, these beliefs will reflect the true statistics of the environment, but individuals can also have inaccurate beliefs about volatility, which can lead to maladaptive learning rates. High learning rates promote forgetting previously learned reward probabilities in favor of patterns in more recent observations. Because learning reward probabilities is implemented differently in AI, there is an associated distinction between learning rates and forgetting rates. Learning rates in AI control how quickly beliefs "solidify" and become resistant to change, which has the effect of reducing exploratory behavior, while forgetting rates control how quickly new observations "over-write" previous learning (and are therefore more similar to learning rates in RL models). To be adaptive, both learning rates and forgetting rates need to match the true statistics of the environment. Otherwise, beliefs may either become rigid/overconfident or beliefs will be unstable/underconfident. Learning/forgetting rates also need not be the same for all types of observations. For example, in previous studies on substance use disorders using active inference models, healthy participants were found to selectively show faster learning rates for losses than substance users, which also predicted changes in symptom severity over time (Smith, Schwartenbeck, et al., 2020; Smith, Taylor, Stewart, et al., 2021). In addition to different learning/forgetting rates for different observations, some related models also assume individuals maintain explicit beliefs about volatility and update those beliefs over time (Mathys et al., 2014)—allowing for changes in the rate of learning/forgetting after each observation.

When considering plausible mechanisms linked to greater SWB, one useful phenotype could correspond to a combination of optimistic prior beliefs and slow forgetting rates (assuming this also matches the statistics of the environment they occupy). This would entail that an individual's optimistic expectations would be resilient to many negative experiences before having strong negative effects on perception and behavior. Another useful phenotype might involve a moderately fast forgetting rate in contexts where one is transitioning from socially hostile to friendly environments—which might prevent pre-existing pessimistic prior beliefs from promoting behavior that could prevent garnering social support.

4.3 Information-seeking and sensitivity to uncertainty

A third relevant aspect of how individuals seek to minimize EFE is through a type of curiosity or exploratory drive—captured by the epistemic value term described in the previous section. One consequence of the equation for EFE is that the value of the expected reward term controls how driven an individual is to seek information vs. reward (Schwartenbeck et al., 2019; Smith, Friston, et al., 2022). Technically, the expected reward term in EFE takes the form of a probability distribution over observations, where observations that are more "probable" are those that are more rewarding (encoded by the parameters within **C** in **Figure 1**). If this "preference distribution" is highly precise (corresponding to one observation having a very high reward value), then reward-seeking will dominate—leading to a type of "risky" behavior in which individuals try to maximize reward before exploring the environment to identify the optimal strategy. Conversely, a very imprecise preference distribution drives a type of "risk-averse" behavior in which individuals engage in excessive information-seeking (i.e., continually trying to reduce uncertainty beyond what is necessary) and take too long to become confident in the best reward-maximizing strategy. In other words, they act as though they are overly concerned that they will make the wrong decision once they begin seeking reward. Empirical studies in AI have only very recently begun to examine differences in preference precision (and hence information-seeking; see (Smith, Schwartenbeck, et al., 2020; Smith, Taylor, Stewart, et al., 2021)). However, work on exploratory drives within expanded RL models has suggested altered patterns of information-seeking in depression, anxiety, and other psychiatric disorders (Aberg et al., 2022; Fan et al., 2021; Smith, Taylor, Wilson, et al., 2021; Waltz et al., 2020). Whether and how information-seeking drives relate to SWB has not been empirically assessed (for some interesting recent theoretical discussion, see (Miller et al., 2022)). Here, one might hypothesize that SWB would be facilitated by a moderate information-seeking drive that supports informed goal-seeking, while preventing individuals from "jumping to conclusions" too early or remaining uncertain after gathering sufficient information.

4.4 Prior beliefs about EFE precision

As described earlier, AI suggests the brain can learn to be more or less confident in its model's ability to use EFE to optimize action (the parameter $\boldsymbol{\gamma}$ in **Figure 1**). In these models, individuals also start with a prior belief about this EFE precision (the parameter $\boldsymbol{\beta}$ in **Figure 1**), which can be estimated in individuals based on task behavior. This can be viewed as a type of metacognition, in that it involves inferring beliefs about the reliability of one's beliefs (Hesp et al., 2021). One study has shown how this relates to dopamine-related brain regions during neuroimaging (Schwartenbeck et al., 2015), and two other studies have linked this measure to differences in self-reported negative affect and decision uncertainty (Smith, Kirlic, Stewart, Touthang, Kuplicki, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, McDermott, et al., 2021). Therefore, one might predict that a stronger prior belief that EFE precision is high should be a mechanism promoting SWB. High EFE precision also leads behavior to be less random and/or less habit-driven (i.e., it down-weights the influence of the habits encoded in the E vector in **Figure 1**)—making behavior more driven by explicit future predictions, which could help account for bidirectional links between SWB and adaptive behavior (Diener et al., 2018). As EFE precision can be understood as a metacognitive belief and has been related to self-reported decision uncertainty, it also has potential links to other work suggesting the importance of metacognitive abilities in promoting SWB (Varshney & Barbey, 2021).

4.5 Planning horizon and decision tree pruning

A further possible dimension of a computational phenotype pertains to prospective planning. This form of planning involves imagining different sequences of actions (policies; encoded by π

in **Figure 1**) and their future outcomes (*o*). One individual difference in this context is planning horizon, which corresponds to how many steps into the future one considers while making decisions (i.e., the length of the action sequence encoded in each possible policy). If this horizon is short, decisions will focus on optimizing short-term outcomes, while a long horizon will optimize long-term outcomes (e.g., even if they require enduring a negative short-term outcome). Another important individual difference corresponds to something called "decision tree pruning" (Huys et al., 2012; Lally et al., 2017), which arises from the fact that it typically takes too much time to imagine the distal future outcomes of all possible plans (i.e., policies). Therefore, a mechanism for focusing on only a few possible plans is needed (metaphorically "pruning away" other branches in a tree of possible diverging paths). This mechanism involves ceasing to simulate/imagine the rest of a possible action sequence once short-term negative outcomes are expected (i.e., one does not "think it through" to consider the possibility of a positive long-term outcome). These two differences—planning horizon and pruning—also represent plausible mechanisms for promoting behaviors that would maintain high vs. low SWB. Both a short planning horizon and too much pruning will lead to myopic planning. For example, this type of planning might prevent an individual from beginning a promising career trajectory because the initial starting position is not enjoyable or respected. Or it could prevent an individual from offering a difficult apology to maintain a fulfilling friendship. Yet, too little pruning is also maladaptive, as decision-making can become inefficient and overwhelming. It is also ideal for planning horizon to be flexible, as distal future outcomes can in some cases be too unpredictable to consider. In other cases, survival can also require fast decisions focused on the immediate future. Thus, one would expect individuals would have higher SWB if they are capable of long planning horizons, but where deployment of this ability is flexible and optimized to the relevant context. Behavioral planning tasks are available that could be used to examine potential relations between these mechanisms and SWB (Huys et al., 2012; Lally et al., 2017), but this remains untested.

4.6 Granular state spaces

The final element of a generative model that we will consider as a potential dimension for computational phenotyping pertains to the specificity of categories used to understand the world (i.e., technically, the number of possible states within the state-space of a model; encoded within *s* in **Figure 1**). Unlike some other model elements discussed above, this is not a single parameter. Instead, it is a more general attribute of the structure of a generative model. One relevant example of this pertains to the specificity of emotion concept learning, and its relation to constructs such as emotion differentiation (Kashdan et al., 2015), emotional complexity/diversity (Kang & Shaver, 2004; Quoidbach et al., 2014), emotional awareness (Lane & Smith, 2021; Lane et al., 2015; Smith, Killgore, & Lane, 2018), and alexithymia (Bagby et al., 1994; Lane et al., 2021; Maroti et al., 2018; Trevisan et al., 2019), all of which share the notion of granularity. Individuals with low emotional granularity tend to use broad, non-specific emotion categories (e.g., "good", "bad"). This can be understood as a generative model with only two possible states or hypotheses (Smith, Lane, et al., 2019; Smith, Parr, et al., 2019), which constrains the amount of information available to guide adaptive choice. In contrast, those with high granularity have many specific emotion concepts (e.g., many types of "bad", such as sad, angry, afraid, guilty, jealous, etc.). This corresponds to a generative model with a large number of possible states that can explain many unique patterns of sensory input and provide precise information to guide effective emotion regulation and adaptive social decision-making (Satpute et al., 2020; Satpute et al., 2016; Smith, Killgore, Alkozei, et al., 2018; Smith, Killgore, & Lane, 2018; Smith & Lane, 2016).

This computational formulation of emotion concept granularity may therefore offer a novel perspective on the beneficial role of emotion knowledge in cognition. For example, it illustrates how a larger and more fine-grained state space for emotion concepts can offer precise predictions to adaptively guide choice. This could also be extended to consideration of unique emotion concepts present in different languages/cultures (Majid, 2012; Russell, 1991; Satpute et al., 2020), and how such concepts could be uniquely helpful in navigating those specific sociocultural contexts. This also supports previous suggestions that adaptive emotional functioning could be increased by acquiring new emotion concepts from other languages/cultures (Barrett, 2017). However, as with other model parameters, more granularity will likely only be useful to the extent that the environment truly has many underlying states. For example, if one grows up in an emotionally impoverished environment where people only express coarse-grained emotional signals, then it may be most adaptive to maintain a model that is not overly complex and assumes there is more variation than is actually present (Smith, Steklis, et al., 2022; Smith, Steklis, et al., 2020). While simulations of such differences have been reported using AI (Smith, Lane, et al., 2019; Smith, Parr, et al., 2019), no behavioral tasks have yet been developed to study the granularity of individuals' generative models. This represents an important future direction, which could potentially draw on existing paradigms used to study cross-cultural differences in other perceptual categorization processes (e.g., color categorization (Twomey et al., 2021; Winawer et al., 2007)). It is also important to note that, while we have used emotion concept specificity as a concrete example, the notion of state-space granularity can be applied to many other relevant domains (e.g., coarse- vs. fine-grained beliefs about personality types, differences in social status, etc.), which will be important to consider in future work as well.

5. Potential empirical applications

As each of the model elements discussed above can differ across individuals, they afford possible explanations for individual differences in perception, learning, and decision-making—each of which could lead to differences in SWB. By extension, these model elements also offer possible explanations for cultural differences, which can be viewed as computational phenotypes that are similar in individuals within a culture but differ from individuals in other cultures. For example, perhaps some cultures place more value than others on reflective practices (e.g., heightened attention to one's own uncertainty, more frequent engagement in prospection/retrospection) that would be expected to increase information-seeking and planning horizon (e.g., see (Fan et al., 2021; Gershman, 2018; Jackson et al., 2020; Kaplan & Friston, 2018; Smith, Taylor, Wilson, et al., 2022)). Or perhaps some cultures tend to share implicit expectations that the social environment is more volatile than others. SWB could also have unique computational correlates in different cultures. Such hypotheses could be tested using standard methods in computational psychiatry to examine significant differences between cultural groups as well as the correlates of within-culture variability (for a summary of possible empirical approaches, see **Table 2** below). For example, if groups of individuals from two distinct cultures were each asked to complete widely-used decision tasks designed to test for differences in information-seeking/preference precision (e.g., (Smith, Schwartenbeck, et al., 2020; Wilson et al., 2014), distal planning (e.g., (Huys et al., 2012)), or volatility beliefs (e.g., (de Berker et al., 2016; Diaconescu et al., 2017; Iglesias et al., 2013; Lawson et al., 2017)), hypotheses could be tested that significant group differences in these traits would be observed. Within each cultural group, possible relationships could also be tested between these traits and SWB. For example, perhaps volatility beliefs are correlated with SWB in some cultures but not in others.

Table 2. Possible empirical approaches for measurement of, and intervention on, different computational model parameters.

Model parameter	**Example experimental paradigms that could be adapted to estimate parameters in contexts relevant to subjective wellbeing**	**Possible intervention approaches**	**Target outcomes of interventions**
Precision of prior beliefs over states (**D** and **B**) and sensory precision (**A**)	Conditioned perception tasks (Powers et al., 2017), interoceptive inference tasks (Smith, Kuplicki, Feinstein, et al., 2020; Smith, Kuplicki, Teed, et al., 2020; Smith, Mayeli, et al., 2021)	Training based on corrective feedback, mindfulness, and/or selective attention (Feinstein et al., 2018; Price et al., 2019; Sugawara et al., 2020; Weng et al., 2021)	A more positive outlook on life due to more optimistic prior beliefs
Learning/forgetting rates & volatility estimates	Perceptual learning tasks (Smith, Mayeli, et al., 2021), multi-arm bandit tasks (Smith, Schwartenbeck, et al., 2020; Smith, Taylor, Stewart, et al., 2021), reversal learning and change-point detecting tasks (Browning et al., 2015; de Berker et al., 2016; Diaconescu et al., 2017; Huang et al., 2017; Iglesias et al., 2013)	Corrective feedback or directed attention-based training (no existing interventions)	Reduced levels of anxiety due to the belief that the world is more predictable
Preference precision (**C**): Reward-seeking vs. information-seeking	Explore-exploit and multi-arm bandit tasks (Fan et al., 2021; Smith, Schwartenbeck, et al., 2020; Smith, Taylor, Stewart, et al., 2021; Wilson et al., 2014)	Increasing awareness of uncertainty (cognitive and behavioral therapeutic interventions; (Barlow et al., 2016; Segal et al., 2004)), altering incentive structure (Ederer & Manso, 2013)	Reduced avoidance behavior, impulsivity, and/or uncertainty due to optimizing the balance between reward-seeking and information-seeking

Model parameter	Example experimental paradigms that could be adapted to estimate parameters in contexts relevant to subjective wellbeing	Possible intervention approaches	Target outcomes of interventions
Expected free energy precision (γ)	Limited offer (risk-taking) task (Schwartenbeck et al., 2015), approach-avoidance conflict task (Smith, Kirlic, Stewart, Touthang, Kuplicki, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, McDermott, et al., 2021)	Interventions targeting self-efficacy, confidence, self-determination, etc.	More confident, value-sensitive behaviors with less influence of unhealthy habits
Planning horizon and decision-tree pruning	Multi-step planning tasks (Huys et al., 2012)	Increasing awareness of the importance of cognitive reflection and prospective planning (cognitive and behavioral therapeutic interventions; (Barlow et al., 2016; Segal et al., 2004)), enforcing delay periods before choice (Bernstein et al., 2018; Shin & Grant, 2021)	A greater ability to work toward healthy long-term goals, despite needing to go through short-term challenges and discomfort
State-space granularity (s)	No existing paradigms. Would require comparison of models with different numbers of hidden states in the context of an emotion inference task	Emotional awareness training interventions (Burger et al., 2016; Farnam et al., 2014; Neumann et al., 2017; Persich et al., 2021; Thakur et al., 2017), promoting identification of alternative interpretations (Barlow et al., 2016; Hendricks et al., 2018)	A more precise understanding of emotions, which is expected to improve emotion regulation and social decision-making

A further consideration is that, because cultural norms and practices are passed on through learning, computational models of learning in both RL and AI can offer hypotheses about this learning process. However, unlike the generic model parameters discussed above, cultural norms and values are content-specific (e.g., that it is acceptable or unacceptable to strive for personal over collective benefit). As such, the approach to testing such hypotheses would instead require *model comparison* (Rigoux et al., 2014). This entails specifying a number of different mechanistic hypotheses—formalized as different generative models—that could explain such differences. For example, one candidate model might specify that people within collectivistic cultures learn a generative model in which self-serving choices are expected to generate non-preferred social feedback, which then deters the selection of those behaviors. In contrast, another candidate generative model might specify that individuals learn preferences for observing prosocial behavior in both themselves and others. In this latter case, behavior would not be driven by expected social feedback, but by the intrinsic desire to observe prosocial behavior. A more complex model might include a competition between the precision of preferences for personal and societal benefit, with a parameter that weights their relative influence. Once a space of possible models is constructed, each model can be fit to available data. One can then perform model comparison to identify which model best accounts for the way this cultural difference is acquired (i.e., which model can best reproduce the experimental data in simulations).

The approaches described in this section can also be applied to many other related phenomena. For example, people in individualistic cultures tend to prefer high-arousal emotions more than those in collectivistic cultures (Lim, 2016). Within AI models there are several possible mechanisms through which these distinct value systems could be acquired during development, each representing a unique hypothesis to be tested. For example, this might involve learning distinct preference distributions (i.e., within **C** in **Figure 1**) over high arousal sensations based on what was experienced most frequently during development, or it could involve associative learning processes (within **A** in **Figure 1**) in which an individual has learned to predict that displays of high-arousal emotions will be met with non-preferred outcomes (e.g., negative social feedback). Or perhaps individuals in collectivistic cultures learn to expect that high-arousal states threaten the successful minimization of VFE and EFE (i.e., that action outcomes are less predictable and that the value of $\boldsymbol{\gamma}$ in **Figure 1** is low).

Another hypothesis is that emotions are conceptualized in distinct ways in different cultures, and that it is this difference in conceptualization that explains differences in emotional experience. In other words, different cultures have different emotion categories that can be inferred to explain their experience, which would be encoded within s in **Figure 1**). This type of emotion concept inference process has been simulated in previous work, but cross-cultural variation has not been addressed (Smith, Lane, et al., 2019; Smith, Parr, et al., 2019). Emotion concept learning in this context can be modeled as learning the probability of making various internal and external observations under different emotional states, $p(o|s)$. For example, observations of threat, high arousal, and avoidance motivation may have high joint probability under the concept of "fear", while observations of pleasant sensations, high arousal, and approach motivations may have high joint probability under the concept of "excitement". However, these probabilistic mappings are learned and can differ by culture (Barrett, 2017; Russell, 1991; Smith, Killgore, & Lane, 2018; Widen & Russell, 2008), which means the same experiences can be interpreted as different emotions in different cultures. In computational terms, this would mean that different cultures have different numbers/types of possible states (i.e., different levels of granularity) and different emotional state-observation mappings encoded in their internal models.

In support of this, different languages often include emotion concept terms that are difficult to translate. For example, the concept of "wabi-sabi" in Japanese roughly corresponds to "appreciating beauty in imperfection" and the concept of "schadenfreude" in German corresponds to "feeling pleasure in response to another person's suffering or misfortune". These concepts do not have 1-to-1 translations to English words. The degree to which this influences emotional experience is an open question, but both active inference and related theories of emotion (e.g., constructivism; (Barrett, 2017)) would predict meaningful influences. As in the previous cases we have described, this could also be tested through model comparison. For example, emotion induction procedures could be combined with behavioral tasks that require decisions to be made based on perceived emotional states. Then models with different numbers/types of emotion concepts could be compared in their ability to account for differences in self-reported emotions and patterns of choice behavior. If different models best explained behavior in different cultural groups, this would support the idea that they use different models as a means of understanding their emotional experience (i.e., different inferences best minimize variational free energy and make different predictions about the choices that will minimize expected free energy).

One additional area of research on SWB that could be opened up using computational approaches is modeling multi-person interactions. In this case, two (or more) individuals cooperate or compete when completing a behavioral task—allowing models to be fit to the behavior of each individual as they react to each other's choices. This can offer the opportunity to evaluate individual differences in parameters of one individual's model *of another person* (i.e., corresponding to the neurocomputational processes underlying mentalization or theory of mind (Amodio & Frith, 2006; Friston & Frith, 2015; Frith & Frith, 2006, 2012; Lombardo et al., 2010; Schurz et al., 2014)). To date, a small number of studies have begun to examine how individuals minimize "social prediction error" in such tasks (Diaconescu et al., 2014; Diaconescu et al., 2017), illustrating sensitivity to beliefs about volatility in the intentions of others, relationships to self-reported empathy, and the potential role of specific neuromodulatory systems, but this work has not focused on individual differences in SWB. Complementary work within the RL framework has also shown that happiness ratings are lower in such tasks when interaction partners are inequitably rewarded (Rutledge et al., 2016), but has not focused on inference processes within internal models of others or the potential role of free energy minimization. Future research could build upon this work by using such tasks in combination with AI models to examine additional hypotheses. For example, perhaps individuals with precise prior beliefs that others tend to be altruistic (and therefore behave in more cooperative, trusting ways) will also be more likely to report higher SWB. Or perhaps individuals that more fully consider the predicted effect of their current choices on the distal future choices of their interaction partner (i.e., greater planning horizon) will also report greater real-world social success (e.g., attainment of greater social support, more meaningful relationships, etc.).

Another complementary research direction could be to use multi-agent simulation to identify public policies that better promote SWB in societies of a given culture (e.g., by formalizing multi-agent interactions in which one agent corresponds to the government and has a particular set of "policy-making" actions to choose from). A branch of game theory called *mechanism design,* which can be thought of as the "engineering" side of economic theory, is one appropriate mathematical approach for identifying policies in this manner (Maskin, 2008). Based on the present discussion, future work might consider mechanism design using agents that are themselves engaged in active inference (e.g., within a general equilibrium macroeconomic model that is being utilized to understand a country's SWB (Hill et al., 2021)). Parallel work in the RL literature has developed similar large-scale simulations to develop dynamic taxation and subsidy policies that consider multiple objectives, policy levers, and behavioral responses from strategic actors that optimize for their individual objectives (Trott et al., 2021). A taxation policy from such reinforcement learning simulations can even outperform optimal static policies in terms of productivity and equity (Zheng et al., 2021). Thus, it would be interesting to see whether the active inference framework could offer any additional benefits in this line of research.

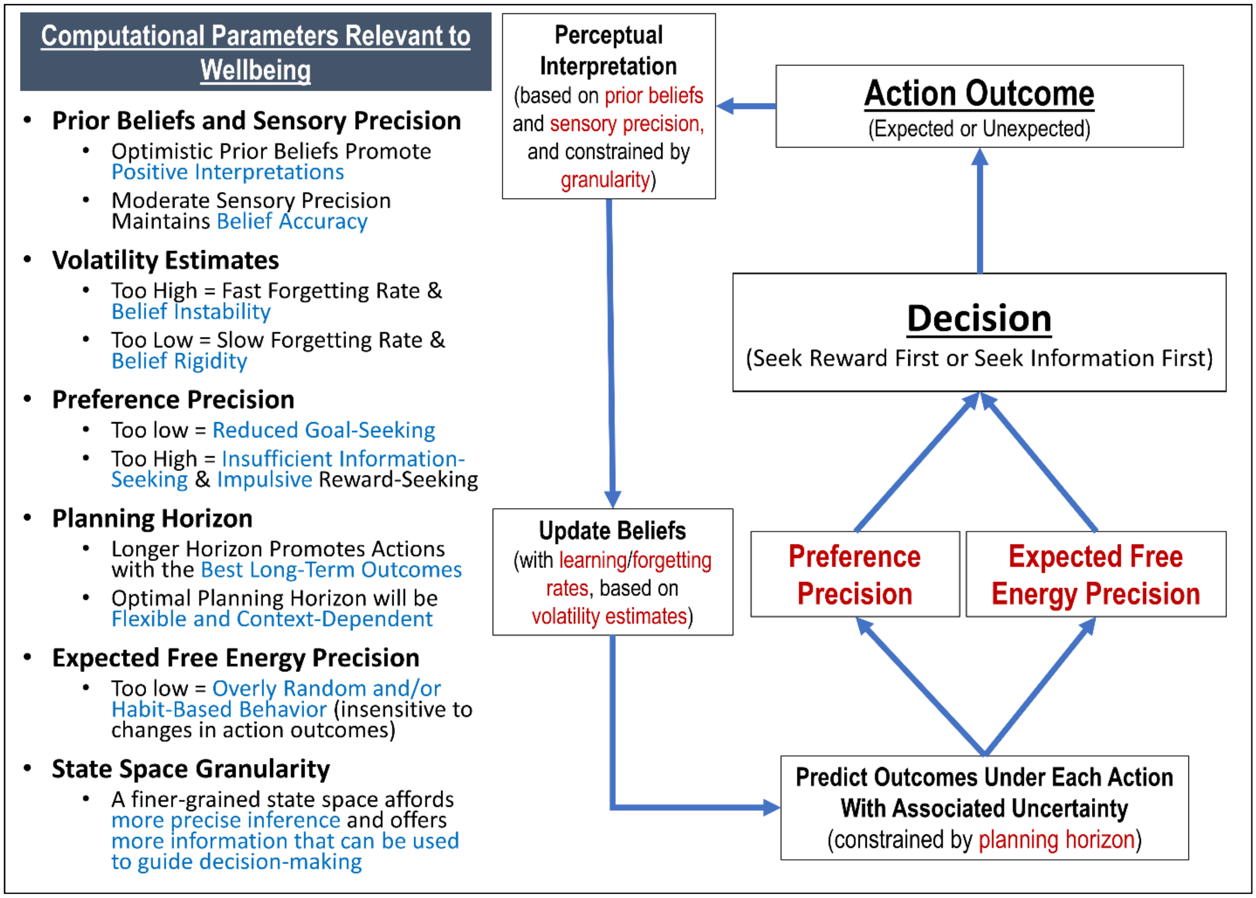
6. Summary and conclusion: Integrated phenotypes for free energy minimization

We have outlined several computational model parameters that can be configured in unique ways in different individuals—representing distinct computational strategies one might use in attempt to minimize free energy. We have also considered how each of these parameters could (alone or in combination) either promote or hinder maintenance of high SWB depending on the values they take and their match to the statistics of the local environment. **Figure 2** (below) provides a visual summary of these parameters and their interactions.

These parameters may link to SWB in at least two ways. First, parameters like EFE precision (and perhaps prior beliefs about states) may have close internal connections with feelings of wellbeing. Second, many other parameters may contribute to wellbeing indirectly through their influence on behavior. That is, if they match well with the local environment, behavior may be more likely to cultivate the career success and social support that contribute to satisfaction with life.

We suggest that the reconceptualization of SWB we have described in terms of levels of success in free energy minimization, and the methods offered by computational modeling to empirically test this new framing, each represent important means of advancing our understanding of SWB. It may offer a novel behavioral approach and set of tools/measures that—while potentially showing convergent validity with standard SWB measures—can also offer novel types of relevant information. It may help to identify specific mechanisms that promote or hinder wellbeing on an individual basis and therefore potentially facilitate the development of personalized interventions aimed at improving wellbeing. We look forward to seeing the potentially important outcomes of this line of future research.

Figure 2. Computational framework for subjective wellbeing.



Computational Parameters Relevant to Wellbeing

- **Prior Beliefs and Sensory Precision**
  - Optimistic Prior Beliefs Promote Positive Interpretations
  - Moderate Sensory Precision Maintains Belief Accuracy
- **Volatility Estimates**
  - Too High = Fast Forgetting Rate & Belief Instability
  - Too Low = Slow Forgetting Rate & Belief Rigidity
- **Preference Precision**
  - Too low = Reduced Goal-Seeking
  - Too High = Insufficient Information-Seeking & Impulsive Reward-Seeking
- **Planning Horizon**
  - Longer Horizon Promotes Actions with the Best Long-Term Outcomes
  - Optimal Planning Horizon will be Flexible and Context-Dependent
- **Expected Free Energy Precision**
  - Too low = Overly Random and/or Habit-Based Behavior (insensitive to changes in action outcomes)
- **State Space Granularity**
  - A finer-grained state space affords more precise inference and offers more information that can be used to guide decision-making

Perceptual Interpretation (based on prior beliefs and sensory precision, and constrained by granularity)

Action Outcome (Expected or Unexpected)

Decision (Seek Reward First or Seek Information First)

Update Beliefs (with learning/forgetting rates, based on volatility estimates)

Preference Precision

Expected Free Energy Precision

Predict Outcomes Under Each Action With Associated Uncertainty (constrained by planning horizon)

Notes. Left: Summary of example parameters and how they could plausibly influence SWB. See main text for further description. Right: Depiction of the circular interactions between perception and action, as well as where the parameters discussed in the text influence learning and decision-making processes contributing to these circular interactions. Specific parameters are highlighted in red font.

Acknowledgements

This work is supported in part by The William K. Warren Foundation, the National Institute of Mental Health (R01MH123691; R01MH127225), the National Institute of General Medical Sciences (P20GM121312), and JSPS KAKENHI (19H01527). The authors would like to thank Nathan Ensz for his help with formatting and proofreading the manuscript.

Conflict of interest statement

None of the authors have any conflicts of interest to disclose.

Authors

Ryan Smith
Laureate Institute for Brain Research; University of Tulsa, Oxley College of Health Sciences
rsmith@laureateinstitute.org

Lav R. Varshney
University of Illinois Urbana-Champaign

Susumu Nagayama
Hitotsubashi University

Masahiro Kazama
Habitech Inc.

Takuya Kitagawa
Wellbeing for Planet Earth Foundation

Yoshiki Ishikawa
Wellbeing for Planet Earth Foundation

Publishing Timeline

Received 2 February 2022
Accepted 12 July 2022
Published 1 November 2022

References

Aberg, K. C., Toren, I., & Paz, R. (2022). A neural and behavioral tradeoff underlies exploratory decisions in normative anxiety. *Molecular Psychiatry*, *27*, 1573–1587. https://doi.org/https://doi.org/10.1038/s41380-021-01363-z

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268-277. https://doi.org/10.1038/nrn1884

Anand, S., & Sen, A. (1992). Human development index: methodology and measurement. In *Human development report office occasional paper no. 12*. UNDP.

Bagby, R. M., Taylor, G. J., & Parker, J. D. (1994). The Twenty-item Toronto Alexithymia Scale--II. Convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research*, *38*(1), 33-40. https://doi.org/10.1016/0022-3999(94)90006-x

Barlow, D. H., Allen, L. B., & Choate, M. L. (2016). Toward a unified treatment for emotional disorders - republished article. *Behavior Therapy*, *47*(6), 838-853. https://doi.org/10.1016/j.beth.2016.11.005

Barrett, L. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society London B: Biological Science*, *371*(1708). https://doi.org/10.1098/rstb.2016.0011

Bernstein, E., Shore, J., & Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(35), 8734-8739. https://doi.org/10.1073/pnas.1802407115

Blain, B., & Rutledge, R. B. (2020). Momentary subjective well-being depends on learning and not reward. *Elife*, *9*. https://doi.org/10.7554/eLife.57977

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment*, *9*, 52–69.

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590-596. https://doi.org/10.1038/nn.3961

Burger, A. J., Lumley, M. A., Carty, J. N., Latsch, D. V., Thakur, E. R., Hyde-Nolan, M. E., . . . Schubiner, H. (2016). 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*, *81*, 1-8. https://doi.org/10.1016/j.jpsychores.2015.12.003

Da Costa, L., Sajid, N., Parr, T., Friston, K. J., & Smith, R. (2020). The relationship between dynamic programming and active inference: the discrete, finite-horizon case. *arXiv*, arXiv:2009.08111.

Danner, D. D., Snowdon, D. A., & Friesen, W. V. (2001). Positive emotions in early life and longevity: findings from the nun study. *Journal of Personality and Social Psychology 80*, 804–813.

de Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, *7*, 10996. https://doi.org/10.1038/ncomms10996

De Neve, J. E., Diener, E., Tay, L., & Xuereb, C. (2013). *World Happiness Report*. UN Sustainable Development Solutions Network.

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., . . . Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618-634. https://doi.org/10.1093/scan/nsw171

Diener, E., & Chan, M. (2011). Happy people live longer: subjective well-being contributes to health and longevity. *Applied Psychology: Health and Well-Being*, *3*, 1-43.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71-75. https://doi.org/10.1207/s15327752jpa4901_13

Diener, E., Oishi, S., & Tay, L. (2018). Advances in subjective well-being research. *Nature Human Behaviour*, *2*(4), 253-260. https://doi.org/10.1038/s41562-018-0307-6

Diener, E., Pressman, S. D., Hunter, J., & Delgadillo-Chase, D. (2017). If, why, and when subjective well-being influences health, and future research needed. *Applied Psychology: Health and Well-Being*, *9*, 133–167.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social indicators research*, *97*(2), 143-156.

Ederer, F., & Manso, G. (2013). Is pay for performance detrimental to innovation? *Management Science*, *59*(7), 1496-1513.

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*, 6149. https://doi.org/10.1038/ncomms7149

Eldar, E., Roth, C., Dayan, P., & Dolan, R. J. (2018). Decodability of reward learning signals predicts mood fluctuations. *Current Biology*, *28*(9), 1433-1439 e1437. https://doi.org/10.1016/j.cub.2018.03.038

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, *20*(1), 15-24. https://doi.org/10.1016/j.tics.2015.07.010

Fan, H., Gershman, S. J., & Phelps, E. A. (2021). Trait somatic anxiety is associated with reduced directed exploration and underestimation of uncertainty. . *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/yx6sb

Farnam, A., Somi, M. H., Farhang, S., Mahdavi, N., & Ali Besharat, M. (2014). The therapeutic effect of adding emotional awareness training to standard medical treatment for irritable bowel syndrome: a randomized clinical trial. *Journal of Psychiatric Practice*, *20*(1), 3-11. https://doi.org/10.1097/01.pra.0000442934.38704.3a

Feinstein, J. S., Khalsa, S. S., Yeh, H., Al Zoubi, O., Arevian, A. C., Wohlrab, C., . . . Paulus, M. P. (2018). The elicitation of relaxation and interoceptive awareness using floatation therapy in individuals with high anxiety sensitivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(6), 555-562. https://doi.org/10.1016/j.bpsc.2018.02.005

Fredrickson, B. L., Tugade, M. M., Waugh, C. E., & Larkin, G. R. (2003). What good are positive emotions in crises? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. *Journal of Personality and Social Psychology*, *84*, 365–376.

Friston, K., & Frith, C. (2015). A duet for one. *Consciousness and Cognition*, *36*, 390-405. https://doi.org/10.1016/j.concog.2014.12.003

Friston, K. J. (2019). A free energy principle for a particular physics. *arXiv* https://doi.org/1906.10184

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry*, *1*(2), 148-158. https://doi.org/10.1016/S2215-0366(14)70275-5

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531-534. https://doi.org/10.1016/j.neuron.2006.05.001

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology 63*, 287-313. https://doi.org/10.1146/annurev-psych-120710-100449

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34-42. https://doi.org/10.1016/j.cognition.2017.12.014

Hartwig, M., Bhat, A., & Peters, A. (2022). How stress can change our deepest preferences: stress habituation explained using the free energy principle. *Frontiers in psychology*, *13*, 865203. https://doi.org/10.3389/fpsyg.2022.865203

Hendricks, R. K., Demjén, Z., Semino, E., & Boroditsky, L. (2018). Emotional implications of metaphor: Consequences of metaphor framing for mindset about cancer. *Metaphor and Symbol*, *33*(4), 267-279.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Computation*, *33*(2), 398-446. https://doi.org/10.1162/neco_a_01341

Hesp, C., Tschantz, A., Millidge, B., Ramstead, M., Friston, K., & Smith, R. (2020). Sophisticated affective inference: simulating anticipatory affective dynamics of imagining future events. In T. Verbelen, P. Lanillos, C. Buckley, & C. De Boom (Eds.), *Active Inference: First International Workshop, IWAI 2020* (Vol. Communications in Computer and Information Science, vol 1326). Springer.

Hill, E., Bardoscia, M., & Turrell, A. (2021). Solving heterogeneous general equilibrium economic models with deep reinforcement learning. *arXiv*. https://doi.org/arXiv:2103.16977 [econ.GN]

Huang, H., Thompson, W., & Paulus, M. P. (2017). Computational dysfunctions in anxiety: failure to differentiate signal from noise. *Biological Psychiatry*, *82*(6), 440-446. https://doi.org/10.1016/j.biopsych.2017.07.007

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*(3), e1002410. https://doi.org/10.1371/journal.pcbi.1002410

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404-413. https://doi.org/10.1038/nn.4238

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*(2), 519-530. https://doi.org/10.1016/j.neuron.2013.09.009

Jackson, B. J., Fatima, G. L., Oh, S., & Gire, D. H. (2020). Many paths to the same goal: balancing exploration and exploitation during probabilistic route planning. *eneuro*, *7*(3). https://doi.org/10.1523/ENEURO.0536-19.2020

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLOS Compututational Biology*, *9*(6), e1003094. https://doi.org/10.1371/journal.pcbi.1003094

Kang, S. M., & Shaver, P. R. (2004). Individual differences in emotional complexity: their psychological implications. *Journal of Personality*, *72*(4), 687-726. https://doi.org/10.1111/j.0022-3506.2004.00277.x

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological Cybernetics*, *112*(4), 323-343. https://doi.org/10.1007/s00422-018-0753-2

Kashdan, T. B., Barrett, L. F., & McKnight, P. E. (2015). Unpacking emotion differentiation: transforming unpleasant experience by perceiving distinctions in negativity. *Current Directions in Psychological Science*, *24*(1), 10-16. https://doi.org/10.1177/0963721414550708

Kim, J., Holte, P., Martela, F., Shanahan, C., Li, Z., Zhang, H., . . . Hicks, J. A. (2022). Experiential appreciation as a pathway to meaning in life. *Nature Human Behaviour*, *6*(5), 677-690. https://doi.org/10.1038/s41562-021-01283-6

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138). https://doi.org/10.1098/rsif.2017.0792

Lally, N., Huys, Q. J. M., Eshel, N., Faulkner, P., Dayan, P., & Roiser, J. P. (2017). The neural basis of aversive Pavlovian guidance during planning. *Journal of Neuroscience*, *37*(42), 10215-10229. https://doi.org/10.1523/JNEUROSCI.0085-17.2017

Lane, R. D., & Smith, R. (2021). Levels of emotional awareness: theory and measurement of a socio-emotional skill. *Journal of Intelligence*, *9*(3). https://doi.org/10.3390/jintelligence9030042

Lane, R. D., Solms, M., Weihs, K. L., Hishaw, A., & Smith, R. (2021). Is the concept of affective agnosia a useful addition to the alexithymia literature? *Neuroscience & Biobehavioral Reviews*, *127*, 747-748. https://doi.org/10.1016/j.neubiorev.2021.05.012

Lane, R. D., Weihs, K. L., Herring, A., Hishaw, A., & Smith, R. (2015). Affective agnosia: expansion of the alexithymia construct and a new opportunity to integrate and extend Freud's legacy. *Neuroscience & Biobehavioral Reviews*, *55*, 594-611. https://doi.org/10.1016/j.neubiorev.2015.06.007

Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, *20*(9), 1293-1299. https://doi.org/10.1038/nn.4615

Lim, N. (2016). Cultural differences in emotion: differences in emotional arousal level between the East and the West. *Integrative Medicine Research*, *5*(2), 105-109. https://doi.org/10.1016/j.imr.2016.03.004

Lomas, T., Case, B., Cratty, F. J., & VanderWheele, T. (2021). A global history of happiness. *International Journal of Wellbeing*, *11*(4), 68-87. https://doi.org/10.5502/ijw.v11i4.1457

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., . . . Baron-Cohen, S. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience*, *22*(7), 1623-1635. https://doi.org/10.1162/jocn.2009.21287

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2003). Reexamining adaptation and the set point model of happiness: reactions to changes in marital status. *Journal of Personality and Social Psychology*, *84*, 527–539.

Luhmann, M., Lucas, R. E., Eid, M., & Diener, E. (2013). The prospective effect of life satisfaction on life events. *Social Psychological and Personality Science*, *4*, 39–45.

Lyubomirsky, S., King, L. A., & Diener, E. (2005). The benefits of frequent positive affect: does happiness lead to success? *Psychological Bulletin*, *131*(6), 803–855.

Majid, A. (2012). The role of language in a science of emotion. *Emotion Review*, *4*(4), 380-381.

Maroti, D., Lilliengren, P., & Bileviciute-Ljungar, I. (2018). The relationship between alexithymia and emotional awareness: a meta-analytic review of the correlation between TAS-20 and LEAS. *Frontiers in Psychology*, *9*, 453. https://doi.org/10.3389/fpsyg.2018.00453

Maskin, E. S. (2008). Mechanism Design: How to Implement Social Goals. *American Economic Review*, *98*(3), 567-576.

Mason, L., Eldar, E., & Rutledge, R. B. (2017). Mood instability and reward dysregulation- a neurocomputational model of bipolar disorder. *Journal of the American Medical Association: Psychiatry*, *74*(12), 1275-1276. https://doi.org/10.1001/jamapsychiatry.2017.3163

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 825. https://doi.org/10.3389/fnhum.2014.00825

Miller, M., Kiverstein, J., & Rietveld, E. (2022). The Predictive Dynamics of Happiness and Well-Being. *Emotion Review*, *14*(1), 15-30.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72-80. https://doi.org/10.1016/j.tics.2011.11.018

Moore, S. M., Diener, E., & Tan, K. (2018). Using multiple methods to more fully understand causal relations: positive affect enhances social relationships. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of Well-Being*. DEF Publishers.

Neumann, D., Malec, J. F., & Hammond, F. M. (2017). Reductions in alexithymia and eotion dysregulation after training emotional self-awareness following traumatic brain injury: a phase I trial. *The Journal of Head Trauma Rehabilitation*, *32*(5), 286-295. https://doi.org/10.1097/HTR.0000000000000277

Pavot, W., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the satisfaction with life scale: evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment*, *57*, 149–161.

Persich, M. R., Smith, R., Cloonan, S. A., Strong, M., & Killgore, W. D. S. (2021). Emotional intelligence training as a protective factor for mental health during the COVID-19 pandemic. *Depression & Anxiety*.

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596-600. https://doi.org/10.1126/science.aan3458

Price, C. J., Thompson, E. A., Crowell, S. E., Pike, K., Cheng, S. C., Parent, S., & Hooven, C. (2019). Immediate effects of interoceptive awareness training through Mindful Awareness in Body-oriented Therapy (MABT) for women in substance use disorder treatment. *Substance Abuse*, *40*(1), 102-115. https://doi.org/10.1080/08897077.2018.1488335

Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, *143*(6), 2057-2066. https://doi.org/10.1037/a0038025

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *Neuroimage*, *84*, 971-985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin*, *110*(3), 426-450. https://doi.org/10.1037/0033-2909.110.3.426

Rutledge, R. B., de Berker, A. O., Espenhahn, S., Dayan, P., & Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*, 11825. https://doi.org/10.1038/ncomms11825

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12252-12257. https://doi.org/10.1073/pnas.1407535111

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2015). Dopaminergic modulation of decision making and subjective well-being. *The Journal of Neuroscience*, *35*(27), 9811-9822. https://doi.org/10.1523/JNEUROSCI.0702-15.2015

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, *52*, 141-166. https://doi.org/10.1146/annurev.psych.52.1.141

Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E., & Moran, R. J. (2019). Locus coeruleus tracking of prediction errors optimises cognitive flexibility: an active inference model. *PLOS Compututational Biology*, *15*(1), e1006267. https://doi.org/10.1371/journal.pcbi.1006267

Sandvik, E., Diener, E., & Seidlitz, L. (1993). Subjective well-being: the convergence and stability of self-report and non-self report measures. *Journal of Personality*, *61*, 317–342.

Satpute, A. B., Nook, E. C., & Cakar, M. E. (2020). The role of language in the construction of emotion and memory: a predictive coding view. In R. D. Lane & L. Nadel (Eds.), *Neuroscience of Enduring Change: Implications for Psychotheray* (pp. 56-88). Oxford University Press.

Satpute, A. B., Nook, E. C., Narayanan, S., Shu, J., Weber, J., & Ochsner, K. N. (2016). Emotions in "black and white" or shades of gray? How we think about emotion shapes our perception and neural representation of emotion. *Psychological Science*, *27*(11), 1428-1442. https://doi.org/10.1177/0956797616661555

Schimmack, U., & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology*, *89*(3), 395–406.

Schneider, L., & Schimmack, U. (2009). Self-informant agreement in well-being ratings: a meta-analysis. *Social indicators research*, *94*, 363–373, Article 363.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9-34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2015). The Dopaminergic Midbrain Encodes the Expected Certainty about Desired Outcomes. *Cerebral Cortex*, *25*(10), 3434-3445. https://doi.org/10.1093/cercor/bhu159

Schwartenbeck, P., & Friston, K. (2016). Computational Phenotyping in Psychiatry: A Worked Example. *eneuro*, *3*(4), ENEURO.0049-0016.2016. https://doi.org/10.1523/ENEURO.0049-16.2016

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, *8*. https://doi.org/10.7554/eLife.41703

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., . . . Ungar, L. (2013). Characterizing geographic variation in well-being using tweets. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 583–591.

Seder, J. P., & Oishi, S. (2012). Intensity of smiling in Facebook photos predicts future life satisfaction. *Social Psychological and Personality Science*, *3*, 407–413.

Mindfulness-Based Cognitive Therapy: Theoretical Rationale and Empirical Status., Mindfulness and acceptance: Expanding the cognitive-behavioral tradition 45-65 (Guilford Press 2004).

Sen, A. K. (1985). *Commodities and Capabilities*. North-Holland.

Shin, J., & Grant, A. M. (2021). When putting work off pays off: The curvilinear relationship between procrastination and creativity. *Academy of Management Journal*, *64*(3), 772–798.

Slavich, G. M., & Irwin, M. R. (2014). From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. *Psychological Bulletin*, *140*(3), 774-815. https://doi.org/10.1037/a0035302

Smith, R., Alkozei, A., Killgore, W. D. S., & Lane, R. D. (2018). Nested positive feedback loops in the maintenance of major depression: an integration and extension of previous models. *Brain, Behavior, and Immunity*, *67*, 374-397. https://doi.org/10.1016/j.bbi.2017.09.011

Smith, R., Friston, K., & Whyte, C. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, (In Press).

Smith, R., Killgore, W. D. S., Alkozei, A., & Lane, R. D. (2018). A neuro-cognitive process model of emotional intelligence. *Biological Psychology*, *139*, 131-151. https://doi.org/10.1016/j.biopsycho.2018.10.012

Smith, R., Killgore, W. D. S., & Lane, R. D. (2018). The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. *Emotion*, *18*(5), 670-692. https://doi.org/10.1037/emo0000376

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., Khalsa, S. S., . . . Aupperle, R. L. (2021). Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: a computational modelling approach. *Journal of Psychiatry & Neuroscience*, *46*(1), E74-E87. https://doi.org/10.1503/jpn.200032

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., McDermott, T. J., . . . Aupperle, R. L. (2021). Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. *Scientific Reports*, *11*(1), 11783. https://doi.org/10.1038/s41598-021-91308-x

Smith, R., Kuplicki, R., Feinstein, J., Forthman, K. L., Stewart, J. L., Paulus, M. P., . . . Khalsa, S. S. (2020). A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLoS Computational Biology*, *16*(12), e1008484. https://doi.org/10.1371/journal.pcbi.1008484

Smith, R., Kuplicki, R., Teed, A., Upshaw, V., & Khalsa, S. S. (2020). Confirmatory Evidence that Healthy Individuals Can Adaptively Adjust Prior Expectations and Interoceptive Precision Estimates. In T. Verbelen, P. Lanillos, C. Buckley, & C. De Boom (Eds.), *Active Inference* (Vol. Communications in Computer and Information Science, vol 1326, pp. 156-164). Springer, Cham. https://doi.org/10.1007/978-3-030-64919-7_16

Smith, R., & Lane, R. D. (2016). Unconscious emotion: a cognitive neuroscientific perspective. *Neuroscience & Biobehavioral Reviews*, *69*, 216-238. https://doi.org/10.1016/j.neubiorev.2016.08.013

Smith, R., Lane, R. D., Parr, T., & Friston, K. J. (2019). Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance. *Neuroscience & Biobehavioral Reviews*, *107*, 473-491. https://doi.org/10.1016/j.neubiorev.2019.09.002

Smith, R., Mayeli, A., Taylor, S., Al Zoubi, O., Naegele, J., & Khalsa, S. S. (2021). Gut inference: A computational modelling approach. *Biological psychology*, *164*, 108152. https://doi.org/10.1016/j.biopsycho.2021.108152

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating Emotions: An Active Inference Model of Emotional State Inference and Emotion Concept Learning. *Frontiers in psychology*, *10*, 2844. https://doi.org/10.3389/fpsyg.2019.02844

Smith, R., Schwartenbeck, P., Stewart, J. L., Kuplicki, R., Ekhtiari, H., Investigators, T., & Paulus, M. P. (2020). Imprecise Action Selection in Substance Use Disorder: Evidence for Active Learning Impairments When Solving the Explore-exploit Dilemma. *Drug and Alcohol Dependence*, *215*, 108208.

Smith, R., Steklis, H. D., Steklis, N., Weihs, K., Allen, J. J. B., & Lane, R. D. (2022). Lower emotional awareness is associated with greater early adversity and faster life history strategy. *Evolutionary Behavioral Sciences*, ebs0000282. https://doi.org/https://doi.org/10.1037/ebs0000282

Smith, R., Steklis, H. D., Steklis, N. G., Weihs, K. L., & Lane, R. D. (2020). The evolution and development of the uniquely human capacity for emotional awareness: A synthesis of comparative anatomical, cognitive, neurocomputational, and evolutionary psychological perspectives. *Biological psychology*, *154*, 107925. https://doi.org/10.1016/j.biopsycho.2020.107925

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., . . . Paulus, M. P. (2021). Slower Learning Rates from Negative Outcomes in Substance Use Disorder over a 1-Year Period and their Potential Predictive Utility. *medRxiv*, 2021.2010.2018.21265152. https://doi.org/10.1101/2021.10.18.21265152

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., . . . Paulus, M. P. (2022). Slower Learning Rates from Negative Outcomes in Substance Use Disorder over a 1-Year Period and Their Potential Predictive Utility. *Computational Psychiatry*, *6*(1), 117-141. https://doi.org/https://doi.org/10.5334/cpsy.85

Smith, R., Taylor, S., Wilson, R. C., Chuning, A. E., Persich, M., Wang, S., & Killgore, W. D. (2021). Lower levels of directed exploration and reflective thinking are associated with greater anxiety and depression. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/3w4je

Smith, R., Taylor, S., Wilson, R. C., Chuning, A. E., Persich, M. R., Wang, S., & Killgore, W. D. S. (2022). Lower Levels of Directed Exploration and Reflective Thinking Are Associated With Greater Anxiety and Depression [Original Research]. *Frontiers in Psychiatry*, *12*. https://doi.org/10.3389/fpsyt.2021.782136

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., . . . Petzschner, F. H. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, *10*, 550. https://doi.org/10.3389/fnhum.2016.00550

Steptoe, A., & Wardle, J. (2011). Positive affect measured using ecological momentary assessment and survival in older men and women. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 18244–18248.

Steptoe, A., Wardle, J., & Marmot, M. (2005). Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes. . *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 6508–6512.

Sugawara, A., Terasawa, Y., Katsunuma, R., & Sekiguchi, A. (2020). Effects of interoceptive training on decision making, anxiety, and somatic symptoms. *BioPsychoSocial Medicine*, *14*, 7. https://doi.org/10.1186/s13030-020-00179-7

Sutton, R., & Barto, A. (1998). *Reinforcement learning: an introduction*. MIT Press.

Tenney, E. R., Poole, J. M., & Diener, E. (2016). Does positivity enhance work performance?: why, when, and what we don't know. *Research in Organizational Behavior*, *36*, 27–36.

Thakur, E. R., Holmes, H. J., Lockhart, N. A., Carty, J. N., Ziadni, M. S., Doherty, H. K., . . . Lumley, M. A. (2017). Emotional awareness and expression training improves irritable bowel syndrome: A randomized controlled trial. *Neurogastroenterol Motil*, *29*(12), e13143. https://doi.org/10.1111/nmo.13143

Trevisan, D. A., Altschuler, M. R., Bagdasarov, A., Carlos, C., Duan, S., Hamo, E., . . . McPartland, J. C. (2019). A meta-analysis on the relationship between interoceptive awareness and alexithymia: distinguishing interoceptive accuracy and sensibility. *Journal of Abnormal Psychology*, *128*(8), 765-776. https://doi.org/10.1037/abn0000454

Trott, A., Srinivasa, S., van der Wal, D., Haneuse, S., & Zheng, S. (2021). Building a foundation for data-driven, interpretable, and robust policy design using the AI economist. *arXiv*. https://doi.org/arXiv:2108.02904

Twomey, C. R., Roberts, G., Brainard, D. H., & Plotkin, J. B. (2021). What we talk about when we talk about colors. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(39). https://doi.org/10.1073/pnas.2109237118

Vanhasbroeck, N., Devos, L., Pessers, S., Kuppens, P., Vanpaemel, W., Moors, A., & Tuerlinckx, F. (2021). Testing a computational model of subjective well-being: a preregistered replication of Rutledge et al. (2014). *Cognition and Emotion*, *35*(4), 822-835. https://doi.org/10.1080/02699931.2021.1891863

Varshney, L. R., & Barbey, A. K. (2021). Beyond IQ: the importance of metacognition for the promotion of global wellbeing. *Journal of Intelligence*, *9*(4). https://doi.org/10.3390/jintelligence9040054

Varshney, L. R., & Varshney, K. R. (2016). Decision making with quantized priors leads to discrimination. *Proceedings of the Institue of Electrical and Electronics Engineers*, *105*(2), 241-255.

Waltz, J. A., Wilson, R. C., Albrecht, M. A., Frank, M. J., & Gold, J. M. (2020). Differential effects of psychotic illness on directed and random exploration. *Computational Psychiatry*, *4*, 18-39. https://doi.org/10.1162/cpsy_a_00027

Weng, H. Y., Feldman, J. L., Leggio, L., Napadow, V., Park, J., & Price, C. J. (2021). Interventions and manipulations of interoception. *Trends in Neurosciences*, *44*(1), 52-62. https://doi.org/10.1016/j.tins.2020.09.010

Weziak-Bialowolska, D., Bialowolski, P., Lee, M. T., Chen, Y., VanderWeele, T. J., & McNeely, E. (2021). Psychometric properties of flourishing scales from a comprehensive well-being assessment. *Frontiers in Psychology*, *12*, 652209. https://doi.org/10.3389/fpsyg.2021.652209

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, *23*(2), 291-312. https://doi.org/10.1016/j.cogdev.2008.01.002

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074-2081. https://doi.org/10.1037/a0038199

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 7780-7785. https://doi.org/10.1073/pnas.0701644104

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2021). AI economist: optimal economic policy design via two-level deep reinforcement learning. *arXiv*. https://doi.org/https://content.apa.org/doi/10.1037/a0038199