Consciousness and Hierarchical Inference

Commentary by Karl Friston (London)

I greatly enjoyed reading Mark Solms's piece on the conscious id. Mindful of writing this commentary, I noted (in the margins of the target article) points of contact between his formulation and our more formal—if somewhat dryer—treatment using Helmholtzian notions of free energy. It was clear, after a few pages, that I was not going to be able to cover every aspect of the remarkable consilience between the two approaches. Instead, I focus on substantiating Solms's key conclusions from the perspective of hierarchical inference in the Bayesian brain.

Keywords: consciousness; free energy; hierarchy; inference; neuronal activity; perception

It strikes me that the neuropsychoanalysis movement—more than any other field—confronts the theoretical challenges that attend affect, emotion, and interoception. While there is an enormous amount of theoretical work on Bayes-optimal perception and motor control in the exteroceptive and proprioceptive domains (Knill & Pouget, 2004; Körding & Wolpert, 2004), there is a curious absence of formal theory pertaining to emotion and interoception. At first glance, one might consider value-learning and optimal decision theory as good candidates for a theory of emotion and affect. However, these normative approaches are rather shallow—appealing tautologically to behaviorist or economic notions such as reward and utility. In what follows, I provide a brief overview of the free-energy principle discussed in the Target Article. This principle provides a framework to revisit the issues of hierarchical representation and conscious and unconscious inference and their location within the cortico–subcortical hierarchy. After considering the neurobiological substrates of conscious inference, I comment briefly on Solms's conclusions about therapeutic interventions.

Karl Friston: The Wellcome Trust Centre for Neuroimaging, University College London, London, U.K.

Acknowledgements: This work was funded by the Wellcome Trust.

Free energy and neurobiology

The free-energy formulation referred to by Solms is an attempt to apply information theory to self-organizing systems like the brain (Friston, 2010). Its premise is simple: to maintain a homoeostatic and enduring exchange with the world (Ashby, 1947), we have to counter perturbations to the states that we expect to be in. In short, we have to minimize surprising violations of our predictions. Mathematically, this surprise cannot be measured directly; however the brain can compute something called *free energy*, which provides a proxy for surprise. Roughly speaking, free energy is prediction error—namely, the mismatch between bottom-up sensations and top-down predictions. These predictions rest on a model of our world that generates predictions in the *exteroceptive*, *proprioceptive*, and *interoceptive* domains. The minimization of exteroceptive prediction error can be cast as perceptual synthesis or inference; the minimization of proprioceptive prediction error corresponds to behavior (as implemented by classical motor reflexes); and the minimization of interoceptive prediction error corresponds to autonomic or visceral homeostasis (mediated by autonomic reflexes). The neuronal substrate of this minimization is probably simpler than one would imagine: a substantial amount of physiological and anatomical evidence suggests that the brain encodes predictions and prediction errors

with the neuronal activity of separable neuronal populations. These two populations pass messages to each other, where populations encoding prediction errors (denoted by ξ in Figure 1) drive populations encoding predictions (denoted by μ in Figure 1). In turn, these predicting populations suppress or inhibit prediction-error populations. Crucially, neuroanatomical evidence suggests that the (generative) model used by the brain is hierarchical, such that top-down predictions try to explain or suppress prediction errors in the level below, while bottom-up prediction errors subvert themselves by informing and optimizing the predictions in the level above (Mumford, 1992; Rao & Ballard, 1999). In this setting, hierarchical predictions come to represent the causes of sensory input in a Bayesian sense. This recurrent and hierarchically deployed process reduces prediction errors at all levels of the hierarchy—thereby optimizing a hierarchical representation (dynamic prediction) of the sensorium, with multiple levels of description.

The analogy between this Helmholtzian suppression of free energy and Freudian free energy is self-evident: the binding of free energy (prediction errors) corresponds to a top-down suppression, which necessarily entails an explanation or resolution of violated predictions. Crucially, the hierarchical structure of generative models—and implicit emergence of nervous energy (activity of prediction error populations)—speaks exactly to Freud's deepest insight, which, states Solms, rests upon the "*depth* (or hierarchy) in the mind."

Hierarchical inference

Clearly, lower levels of hierarchical inference are closer to the sensorium and represent more elemental (and transient) causes of sensory input. Conversely, higher levels of the hierarchy can "see" multiple input modalities. At this point, we start to see the structural basis of Solms's dichotomy between the *autonomic* body (representations or predictions of interoceptive input) and the *somatomotor* body (representations of exteroceptive and proprioceptive input), where these domains converge at higher levels (see Figure 1). Put another way, high-level intransigent representations (mental solids) have an amodal aspect and provide bilateral top-down interoceptive and exteroceptive predictions. In this sense, high-level representations have, necessarily, interoceptive attributes. This resonates with the notion that high-level (e.g., executive or second-order) representations are supported by—or derived from—interoceptive representations. It also suggests that affect is an intrinsic property of the brain: Solms states that "Affect may accordingly be described as an interoceptive sensory modality—but that is not all it is."

In terms of hierarchical inference, affect is a construct or attribute of a higher level representation that is used to explain interoceptive inputs at a lower level—in the same sense that color is used to explain wavelength-selective responses in early visual cortex (Zeki & Shipp, 1988). However, the parallel between hierarchical inference and the dichotomy developed by Solms rests upon a mapping between inference and consciousness.

Free energy and consciousness

The original writings of Helmholtz (1866) focused on unconscious inference in the visual domain. However, in hierarchical (deep) inference schemes (Dayan, Hinton, & Neal, 1995), it is tempting to associate probabilistic representations—encoded by the activity of populations encoding predictions—with consciousness. Many of the attributes of consciousness are shared with these probabilistic representations. In brief, these probability distributions (known as *posterior beliefs*) are encoded by their *sufficient statistics*, such as their mean and variance. For example, the posterior mean or *expectation* is encoded by the activity of populations encoding predictions. This is important because it means that a probabilistic representation is induced by biophysical states of the brain—and uniquely associates one representation (consciousness) with one brain. However, the representation is not the biophysical state that induces it—in the sense that a probability distribution is not the same as its mean and variance. Intuitively, this means that I cannot possess your beliefs (consciousness), but I can believe you believe (I can have beliefs about your biophysical states). If one admits a mapping between consciousness and the probability distribution induced by expectations or predictions, then the hierarchical architecture of our brains has profound implications for consciousness and the arguments pursued by Solms.

Where is the top (center) of the hierarchy?

A tenet of Solms's argument is his deconstruction of the corticocentric view of consciousness. He argues (with compelling empirical evidence) that consciousness resides in (or is generated from) upper-brainstem structures, which may be embellished by (or support) cortical elaborations. In what sense is this consistent with hierarchical inference in the brain? Hierarchical

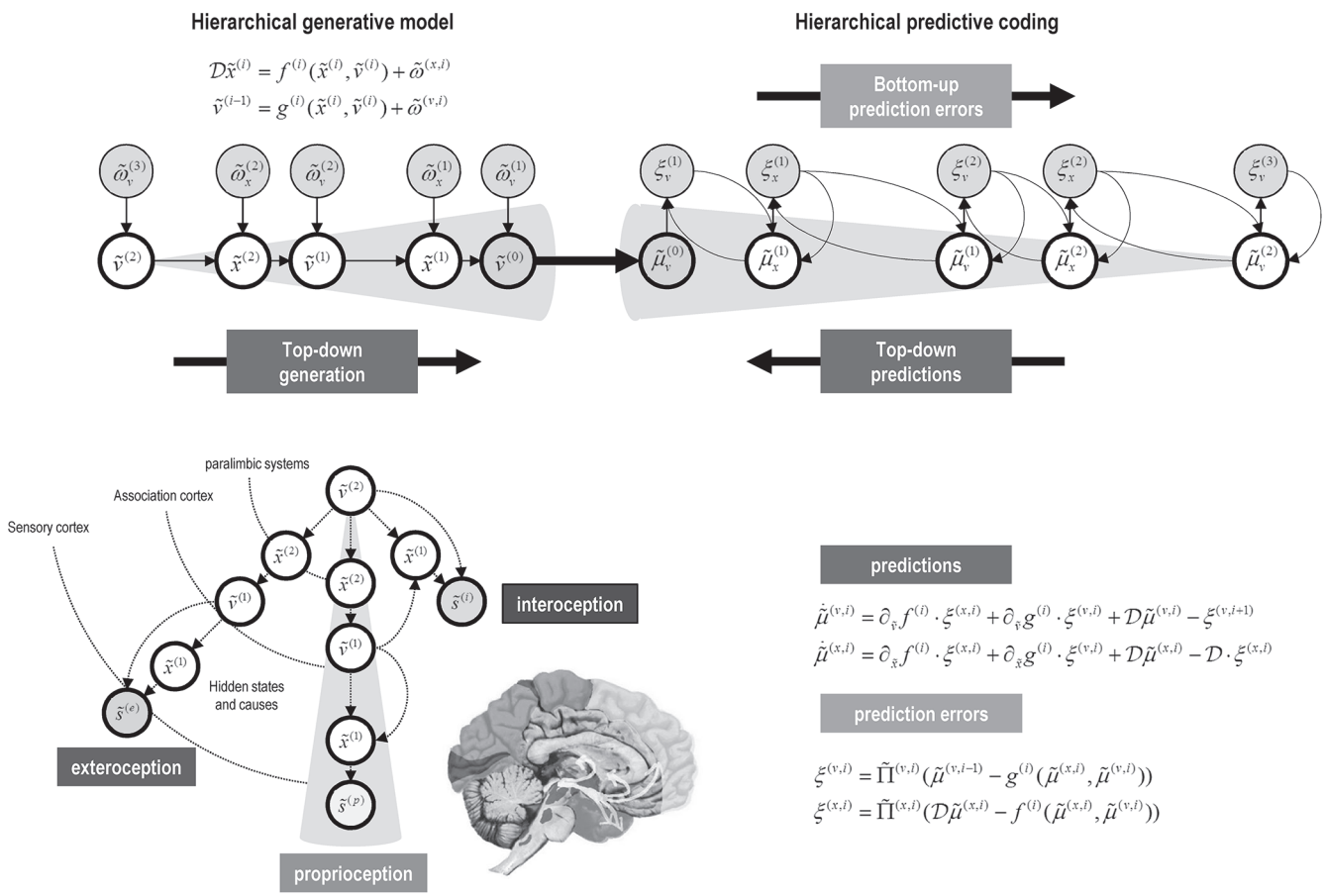

Figure 1. The putative neuronal architectures that might optimize posterior beliefs about the state of the world, using hierarchical generative models. Upper panel: part of a generative model is shown on the left, in terms of a cascade of hidden states and causes in the world that produce sensory input. This architecture is mirrored by hierarchies in the brain that try to explain the input and (implicitly) come to represent the hidden states. Although the details are not important, hidden states $\tilde{x}^{(i)}$ model dynamic dependencies in the way that sensory information is generated, while hidden causes $\tilde{v}^{(i)}$ link hierarchical levels and provide the generative model with a deep structure. The stochastic differential equations (upper left) provide a mathematical specification of the model. The brain infers the hidden causes of sensory input, in terms of posterior expectations or predictions about the hidden states and causes at each level of the hierarchy. The most popular scheme for this hierarchal inference is known as predictive coding, illustrated in the upper right. In this scheme, hidden causes and states are represented in terms of predictions $(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)})$ that are driven by prediction errors $(\xi_x^{(i)}, \xi_v^{(i)})$. Crucially, prediction errors are passed forward to provide bottom-up guidance to neuronal populations encoding predictions, while top-down predictions are assembled to form prediction errors. This process continues until prediction error has been minimized and the predictions become optimal in a Bayesian sense. Lower left panel: this illustrates a model generating exteroceptive, proprioceptive, and interoceptive sensations. Again, the details of this model are not important—it is just meant to illustrate that hierarchical models can have a form in which there is no top but, rather, a center. In this schematic, the scale of grey corresponds to the insert (medial view of the brain): medium grey denotes primary sensory (exteroceptive) input; light grey denotes proprioceptive input; and dark grey denotes interoceptive input. This scheme is based on Figure 1 in the Target Article. The key thing to take from this schematic is that interoceptive parts of the hierarchy are more intimately associated with the center, relative to the peripheral (primary sensory) cortex. Lower right panel: these are the equations that minimize prediction error or free energy (using a gradient descent). They are provided to indicate that the free-energy principle prescribes specific and biologically plausible neuronal dynamics. It can be seen that these equations are based on quantities specified by the generative model and have a relatively simple mathematical form. Of particular note is that the prediction errors are scaled by precision—denoted by $\tilde{\Pi}^{(i)}$. For details about mathematical form and notation, see Friston (2008).

probabilistic representations exist at all levels of the hierarchy. In this context, Solms's arguments make perfect sense, in that different representational attributes can be associated with different locations within the hierarchy. For example, representations with an affective aspect (the id) could be located in systems making interoceptive predictions and coexist (necessarily) with somatomotor representations in the cortex. However, this does not address the question of which "is intrinsically conscious." Let us assume that *intrinsically conscious* means hierarchically supraordinate, in the sense that intrinsic predictions entail extrinsic (exteroceptive and proprioceptive) predictions. So what is the evidence that brainstem regions and associated (para-)limbic brain systems are hierarchically supraordinate to cortical systems? There are two simple lines of evidence—one obvious and one not. It is obvious that primary sensory cortex is at the lowest level of the hierarchy. In other words, there are representations at the cortical level that are only a few synapses away from the sensorium. Clearly, higher order representations of "things and words" that have a temporal persistence involve association cortex. However, to treat the cortex as a functionally homogeneous epicenter is untenable—indeed, it is often depicted on the periphery of centrifugal hierarchies (see Figure 1; see also Mesulam, 1998). The second—less obvious—reason appeals to the inference framework above. In hierarchical inference, top-down predictions fulfill the role of something called *empirical priors* (predictions about predictions). However, at the top (or center) of the hierarchy there are no top-down predictions, and expectations become *full priors*. These expectations are usually associated with the instincts and prior beliefs about bodily states that are selected by evolution (necessary for survival). Neuroanatomically, instinctual or innate priors are concerned with interoceptive inputs and may be entailed by the circuitry and physiology of the upper-brainstem, limbic, and paralimbic systems. This is entirely consistent with the intrinsic representations ascribed to these areas. As discussed in the next section, there is one further reason why these particular systems have an important role in specifying prior (instinctual) beliefs, which touches on the implications for therapy.

Precision, uncertainty, and therapeutic efficacy

Crucially, top-down predictions are not just about the content of lower level representations but also predict their reliability or *precision* (denoted by $\tilde{\Pi}$ in Figure 1). Mathematically, precision is inverse variance or uncertainty. This sort of top-down prediction is thought to be mediated by neuromodulatory mechanisms that optimize the (attentional) gain of populations encoding prediction errors (Feldman & Friston, 2010). This is sensible, in that boosting precise prediction errors gives them a preferential or selective influence on higher (deeper) hierarchical inference. The key thing here is that the precision has itself to be predicted. This means that particular brain systems broadcast posterior beliefs about the precision of various interoceptive and proprioceptive representations—and can, effectively, choose what to explain.

The Bayes-optimal encoding of precision in the brain has already been discussed in terms of attention and affordance and even as an explanation for the emergence of hysterical symptoms (Edwards, Adams, Brown, Pareés, & Friston, 2012). Furthermore, it may provide an interesting metaphor for the repression of (Freudian) free energy, through the neuromodulatory suppression of prediction error units encoding (Helmholtzian) free energy. In the present context, predictions about where precision should be deployed within a hierarchy may be encoded by the activity of classical neuromodulatory transmitter systems that ascend from the extended reticular activating system and upper brainstem—identified in the Target Article. Indeed, at a most basic level, it is this system that controls (through neuromodulatory efferents) the basic cycles of conscious level associated with sleeping and waking (Hobson, 2009). Another classical example is dopamine, which has not only been implicated in neuropsychiatric disorders such as schizophrenia and Parkinsonism but plays a central role in theories of value-dependent learning and emotional behavior (Schultz, 1998). In short, the interior of the brain houses not only the systems necessary for consciousness but also elaborates some of the most important top-down predictions that set the tone for inference elsewhere in the brain—namely, predictions about the precision or salience of prediction errors in one modality (or level) in relation to another.

Therapeutically, as intimated by Solms, locating an intrinsically conscious (representational) capacity at the subcortical and paralimbic level may have important implications for therapy. The nice thing here is that viewing the brain as an inference machine (Dayan, Hinton, & Neal, 1995) means that one can easily motivate therapeutic interactions in terms of changing (posterior) beliefs. At the same time, one can understand this optimization in the context of how we represent precision or uncertainty and the role of key neurotransmitter systems such as the dopaminergic system. In one sense, the therapeutic relationship may

provide, as Solms states, the (unattainable) state of Nirvana "that we now learn, to our surprise, is what the ego aspires to."

REFERENCES

Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *Journal of General Psychology*, 37: 125–128.

Dayan, P., Hinton, G. E., & Neal, R. (1995). The Helmholtz machine. *Neural Computation*, 7: 889–904.

Edwards, M. J., Adams, R. A., Brown, H., Pareés, I., & Friston, K. J. (2012). A Bayesian account of "hysteria." *Brain,* 135 (1): 3495–3512.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4: 215.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4 (11), e1000211.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience*, 11 (2): 127–138.

Helmholtz, H. (1866). Concerning the perceptions in general. In: *Treatise on Physiological Optics, Vol. 3* (3rd edition), trans. J. Southall. New York: Dover, 1962.

Hobson, J. A. (2009). REM sleep and dreaming: Towards a theory of protoconsciousness. *Nature Reviews. Neuroscience*, 10 (11): 803–813.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience*, 27 (12): 712–719.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427 (6971): 244–247.

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, 121: 1013–1052.

Mumford, D. (1992). On the computational architecture of the neocortex. II. *Biological Cybernetics*, 66: 241–251.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2 (1): 79–87.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80 (1): 1–27.

Zeki, S., & Shipp, S. (1988). The functional logic of cortical connections. *Nature*, 335: 311–317.