Brain and Cognition 112 (2017) 64–68

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

The role of interoceptive inference in theory of mind

Sasha Ondobaka *, James Kilner, Karl Friston

Wellcome Trust Centre for Neuroimaging, University College London, United Kingdom
Sobell Department for Motor Neuroscience and Movement Disorders, University College London, United Kingdom

ARTICLE INFO

Article history:
Received 2 May 2015
Revised 3 August 2015
Accepted 4 August 2015
Available online 12 August 2015

Keywords:
Free energy
Predictive processing
Emotion
Intention
Expectation

ABSTRACT

Inferring the intentions and beliefs of another is an ability that is fundamental for social and affiliative interactions. A substantial amount of empirical evidence suggests that making sense of another's intentional and belief states (i.e. theory of mind) relies on exteroceptive (e.g. visual and auditory) and proprioceptive (i.e. motor) signals. Yet, despite its pivotal role in the guidance of behaviour, the role of the observer's interoceptive (visceral) processing in understanding another's internal states remains unexplored. Predicting and keeping track of interoceptive bodily states – which inform intentions and beliefs that guide behaviour – is one of the fundamental purposes of the human brain. In this paper, we will focus on the role of interoceptive predictions, prescribed by the free energy principle, in making sense of internal states that cause another's behaviour. We will discuss how multimodal expectations induced at deep (high) hierarchical levels – that necessarily entail interoceptive predictions – contribute to inference about others that is at the heart of theory of mind.

1. Understanding others' minds

1.1. Mechanisms for inferring others' minds

Understanding or inferring of another's intentions, feelings and beliefs is a hallmark of human social cognition often referred to as mentalising or having a Theory of Mind (ToM; Frith & Frith, 1999; Saxe & Kanwisher, 2003). ToM has been described as a cognitive ability to infer the mental states (intentions and beliefs) of others, through processing of their physical appearance and overt behaviour (e.g. clothes, bodily and facial expressions). Typically, neuronal computations underlying ToM have been associated with multimodal brain regions like the superior temporal sulcus (STS), temporoparietal junction (TPJ) and medial frontal cortex (MFC; Frith & Frith, 2006; Saxe, 2006). However, the nature of processing in higher multimodal regions that accumulates information from different streams remains poorly understood.

1.2. Predictive mechanisms in theory of mind

Recent views propose that predictive mechanisms could play a role in ToM (Hohwy, 2013; Kilner & Frith, 2008; Koster-Hale & Saxe, 2013). In brief, hypotheses about the intentions of others are tested against their observed behaviour by generating top-down predictions of that behaviour – and updating competing hypotheses on the basis of ensuing prediction error. Crucially, the repertoire of hypotheses that can be entertained is borrowed from the constructs (hypotheses) that cause one's own behaviour. This provides a nice explanation for the role of the "mirroring mechanism" (Rizzolatti & Sinigaglia, 2010) in action observation and theory of mind – and the integration of multimodal data inherent in descending multimodal predictions (Kilner & Frith, 2008; Ondobaka & Bekkering, 2013; Ondobaka, de Lange, Wittmann, Frith, & Bekkering, 2014).

However, this perspective only addresses the predicted consequences of movement and says little about the predicted consequences of internal bodily states that contextualise behaviour. In other words, it is unclear how processing of internal visceral/autonomic information (interoception) could contribute to the understanding of others' intentions. There are two ways of thinking about the role of interoception in ToM. The first relates to how processing of exteroceptive information about another's interoceptive state helps us to infer states of mind that cause their behaviour. The second rests on how knowing the interoceptive causes of our own behaviour helps us predict and infer another's. In this paper, we consider interoceptive inference as a special case of active inference, under the free energy principle (Friston, 2010) – and emphasise its potentially fundamental role in grounding the process of inferring another's state of mind from their perceived (motor and autonomic) behaviour (Fig. 1).

* Corresponding author at: Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, United Kingdom.
E-mail address: s.ondobaka@ucl.ac.uk (S. Ondobaka).

http://dx.doi.org/10.1016/j.bandc.2015.08.002
0278-2626/© 2015 The Authors. Published by Elsevier Inc.

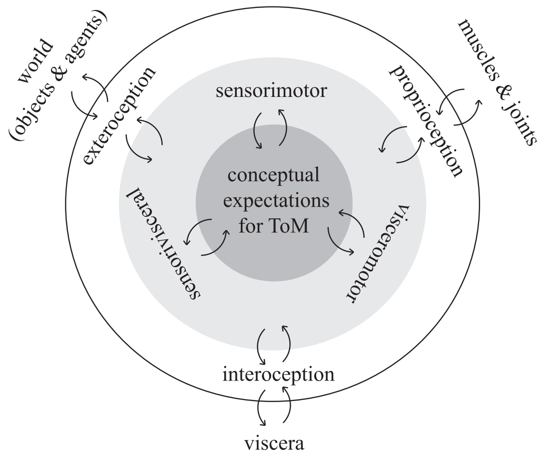

Fig. 1. Schematic representation of a hierarchical predictive neural model for ToM that includes interoception, exteroception and proprioception. White-to-dark grey colour scale represents the neural hierarchy, in which conceptual expectations (dark grey) that include interoception sit high (deep) in the hierarchy. Arrows indicate hierarchical message passing in forward and backward directions carrying prediction error and expectation/prediction, respectively.

2. The free energy and interoceptive inference

2.1. The free energy principle

The free energy principle requires the brain to generate continuous predictions in order to achieve its goal of minimizing free energy – an information theoretical quantity that reflects surprise or prediction error (Friston, 2009). Prediction errors are the difference between sensations and predictions of those sensations based upon an internal or generative model. The basic idea is that the brain constitutes a hierarchical generative model, through which the wealth of incoming information from the viscera, musculoskeletal system and the outside world is interpreted. This interpretation corresponds to inferring the causes of sensations in terms of representations or expectations that would generate the same sensory information, under the hierarchical model. Inferring the causes of visceral/autonomic, motor and sensory (e.g. visual, auditory) information corresponds to interoception, proprioception and exteroception, respectively. It is assumed that all the sensory streams in the brain are organised in a hierarchical fashion, with areas that sit higher (deeper) in the hierarchy representing more abstract information and generating expectations of lower levels (Ondobaka & Bekkering, 2013). At the apex of unimodal sensory hierarchies, multimodal brain regions encode conceptual expectations that necessarily accumulate unimodal information – or prediction errors (Fig. 1; Barrett, 2014; Ondobaka et al., 2014).

2.2. Two modes of free energy minimisation

Minimisation of free energy (prediction error) can occur in two distinct modes, we can either change our predictions to match sensations or we sample sensations to match predictions. Predictive coding typically refers to changing our predictions to match sensations (Friston, 2005; Rao & Ballard, 1999). Controlling motor and autonomic (visceral) system to experience sensations that match our predictions is known as active inference (Friston, Mattout, & Kilner, 2011; Joffily & Coricelli, 2013). These complementary modes of minimising (exteroceptive and proprioceptive/interoceptive) prediction errors correspond to what we generally view as perception or inference and action or motor/autonomic control respectively. It is crucial to note here that two modes of free energy minimisation exist in both proprioceptive and interoceptive domains.

Perception minimises free energy by concurrent dynamical updating of expectations about the causes of external (exteroceptive) and internal (interoceptive and proprioceptive) sensory inputs. For example, perception of a surprising object is associated with an attempt to suppress visual prediction error (Rao & Ballard, 1999). Action, on the other hand, minimises prediction error by directly altering sensory inputs through movement and visceral control that fulfil proprioceptive and interoceptive predictions. For example, movement of the arm is driven by classical motor reflects arcs in the spinal-cord to suppress proprioceptive prediction error – such that descending proprioceptive predictions become motor commands that are reflexively executed by striated muscles (Adams, Shipp, & Friston, 2013). Similarly, the intensity and frequency of on-going contractions of the heart muscle can be modulated to suppress the interoceptive prediction error signalling surprising interoceptive states related to e.g. blood pressure (e.g. Kumagai et al., 2012). This reflexive suppression of interoceptive prediction error corresponds to autonomic reflexes mediated by smooth muscles.

2.3. Free energy and experience of intention and emotion

Proprioceptive and interoceptive prediction errors (free energy) used in motor and autonomic control might relate to our experience of intention and emotion (Seth, 2013; Shipp, Adams, & Friston, 2013). Recent accounts of interoceptive inference have proposed that emotion could be understood from the perspective of hierarchical interoceptive inference (Joffily & Coricelli, 2013; Seth, 2013). For example, Seth (2013) views emotional content as the product of active inference about the likely internal and external causes of visceral changes. Joffily and Coricelli (2013) associated the rate of change of interoceptive prediction error with emotional valance, such that a shift from less expected/valued state (i.e. high free energy) to a more expected/valued state (i.e. low free energy) leads to positive valence. Conversely, negative valance corresponds to a shift from a low free energy to a high free energy state. Although the primary drive for motor and autonomic (visceral) control are descending proprioceptive and interoceptive predictions, these predictions are contextualised by deep hierarchical models that are necessarily accountable to conceptual representations that also generate exteroceptive predictions (Barrett, 2014; Barrett & Simmons, 2015). This enables the integration of exteroceptive information to contextualise (adaptive) motor and visceral responses. For example, when my interoceptive prediction errors signal hunger, I intend (expect) to move my arm to open the fridge because this is what I normally do when I feel (infer myself to be) hungry and find myself in the kitchen. We assume that these deep multimodal levels of representation that guide own behaviour play an important role in inference of others' intentions and emotion.

3. Inferring another's intentions and emotions

3.1. Interoceptive inference mechanism

Interoception refers to one of the fundamental purposes of the human brain – to predict and maintain internal bodily states within physiological bounds and relatively constant over time (Craig, 2009; Critchley & Harrison, 2013). Interoception or interoceptive inference can be viewed as a generalisation of active inference (Clark, 2013; Friston, 2010) to the processing of interoceptive signals carrying information about visceral states (e.g. heart rate, blood pressure, temperature). For example, recent computational work by Joffily and Coricelli (2013) suggests that rate of change of interoceptive free energy corresponds to experience of emotional valance and dynamical adaptation of behaviour. The interoceptive (visceral) processing hierarchy comprises of the brainstem