Consciousness and Cognition 25 (2014) 67–76

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Review

Bayesian inferences about the self (and others): A review

Michael Moutoussis [a,*], Pasco Fearon [b], Wael El-Deredy [c], Raymond J. Dolan [a], Karl J. Friston [a]

[a] Wellcome Trust Centre for Neuroimaging, UCL, United Kingdom
[b] Department of Psychology, University College London, United Kingdom
[c] School of Psychological Sciences, University of Manchester, United Kingdom

ARTICLE INFO

Article history:
Received 28 May 2013
Available online 25 February 2014

Keywords:
Self-representation
Other-representation
Free energy minimisation
Active inference
Personality disorder
Paranoia

ABSTRACT

Viewing the brain as an organ of approximate Bayesian inference can help us understand how it represents the self. We suggest that inferred representations of the self have a normative function: to predict and optimise the likely outcomes of social interactions. Technically, we cast this predict-and-optimise as maximising the chance of favourable outcomes through active inference. Here the utility of outcomes can be conceptualised as prior beliefs about final states. Actions based on interpersonal representations can therefore be understood as minimising surprise – under the prior belief that one will end up in states with high utility. Interpersonal representations thus serve to render interactions more predictable, while the affective valence of interpersonal inference renders self-perception evaluative. Distortions of self-representation contribute to major psychiatric disorders such as depression, personality disorder and paranoia. The approach we review may therefore operationalise the study of interpersonal representations in pathological states.

Contents

* Corresponding author. Address: Wellcome Trust Centre for Neuroimaging at UCL, 12 Queen Square, London WC1N 3BG, United Kingdom.
E-mail address: fzsemmo@gn.apc.org (M. Moutoussis).

http://dx.doi.org/10.1016/j.concog.2014.01.009
1053-8100/

1. Introduction: Agency and the interpersonal self

The sense of self may be experienced at many levels – from the elementary, pre-verbal 'minimal self' that accompanies all conscious perception, to the purposeful, historically constructed 'narrative' self who takes action under the conscious guidance of goal and context. Research based on the idea of the brain as a probabilistic inference device has seen great advances in recent years (Chater & Oaksford, 2008; Friston & Stephan, 2007), allowing important aspects of the minimal and narrative self during perception and action to be considered in the light of how probabilistic prediction interacts with sensory evidence. The computations that brains perform to predict and hypothesis-test underlie what it is like to be an *I who expects* the consequences of acting and perceiving – now and through time (Hohwy, 2007). In this article, we extend this work to self-perception in the interpersonal domain, while acknowledging that the sense of self is important, even in the absence of interactions with others. Simple observation suggests that the interpersonal self is as complicated in its detailed mechanics as it is blatant about its presence. While the minimal self at the core of near-instantaneous perception is difficult to put into words, there is, in the first instance, nothing difficult about putting the interpersonal self into words: 'am a kind person', 'am not as good as her'. The English language describes this powerful self-perception with expressions such as 'he is a terribly self-conscious'.

We claim that the interpersonal self is *actively inferred* during social exchanges and that many of its properties *correspond to the means and ends of a machinery of probabilistic inference*. Inferring self-representations may thus *help achieve desirable ends* (social outcomes). The evidence that we marshal to develop this argument comes from a wide variety of sources, including computational neuroscience, brain imaging, psychiatry, social and clinical psychology. We develop our claim as follows. First, we provide evidence that some high-level, affectively coloured (pain) perception is well described in terms of basic Bayesian reasoning. Second, we describe an extended framework of approximate Bayesian reasoning, namely active inference, which encompasses agency and decision-making. Third, we review the kinds of psychological construct upon which active inference may operate – and show that inferring about these constructs subserves important goals. Fourth, we suggest a model of interpersonal exchange that could form the basis for empirical study. Finally, we examine the psychiatric relevance of making affectively charged inferences, especially about the self. We conclude by discussing the limitations of our approach and the implications for future research.

2. Simple Bayesian inference in high-level perception

2.1. Using Bayesian inference to make sense of experience

The Bayesian approach considers probabilities to be degrees of belief, so that Bayesian inference has the following form. If I make an observation o, what should become of my belief $P(S = s)$ that some relevant aspect of the world is in state s? For example,[1] if o = 'Emil gave me a present', what should become of my belief 'I am a bad person'? If the new observation is surprising – with respect to the existing belief framework – the framework is poor at predicting the observation. It therefore needs to be updated if it is to describe the world more adequately. This updating of beliefs is the essence of Bayesian inference, which adjusts the agent's model of the world so as to render new observations (data) less unpredictable. Although a full description of this well-established formal approach is outside the scope of the present article, the interested reader is referred to (Chater & Oaksford, 2008; Friston & Stephan, 2007; Friston et al., 2013; King-Casas et al., 2008). The claim we make in this paper is that this inferential framework applies to all beliefs – including beliefs about the self.

In a Bayesian framework *what the brain minimises as it makes inferences, including inferences about the self, is unpredictability* and not, for example, proximal discomfort. We will consider an example of this below, in the case of perception of pain. We reformulate the principle of psychological economy as follows: the *primary gain* of a representation is its power to predict outcomes that matter under some prior beliefs. Maximising predictability is equivalent to minimising surprise. Clearly, surprising outcomes rest upon prior beliefs. In our case, these beliefs will be about the self (and others). Crucially, surprise can be quantified as the negative log (Bayesian) evidence for a model. This means that minimising surprise maximises the evidence for a model or representation of interpersonal exchange.

[1] This is a real example, as will be discussed in the section on clinical implications of our proposal.