ARTICLE IN PRESS

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Review

The free energy principle for action and perception: A mathematical review

Christopher L. Buckley [a,*,1], Chang Sub Kim [b,1], Simon McGregor [a], Anil K. Seth [a,c]

[a] School of Engineering and Informatics, Evolutionary and Adaptive Systems Group, University of Sussex, Brighton, BN1 9QJ, UK
[b] Department of Physics, Chonnam National University, Gwangju 61186, Republic of Korea
[c] Sackler Centre for Consciousness Science, University of Sussex, Brighton, BN1 9QJ, UK

HIGHLIGHTS

- The free energy principle (FEP) is suggested to provide a unified theory of the brain, integrating data and theory relating to action, perception, and learning.
- We provide a complete mathematical guide to a suggested biologically plausible implementation of the FEP.
- A simple agent-based model implementing perception and action under the FEP.

ARTICLE INFO

Article history:
Received 27 April 2017
Received in revised form 2 August 2017
Available online xxxx

Keywords:
Free energy principle
Perception
Action
Inference
Bayesian brain
Agent-based model

ABSTRACT

The 'free energy principle' (FEP) has been suggested to provide a unified theory of the brain, integrating data and theory relating to action, perception, and learning. The theory and implementation of the FEP combines insights from Helmholtzian 'perception as inference', machine learning theory, and statistical thermodynamics. Here, we provide a detailed mathematical evaluation of a suggested biologically plausible implementation of the FEP that has been widely used to develop the theory. Our objectives are (i) to describe within a single article the mathematical structure of this implementation of the FEP; (ii) provide a simple but complete agent-based model utilising the FEP and (iii) to disclose the assumption structure of this implementation of the FEP to help elucidate its significance for the brain sciences.

Contents

* Corresponding author.
E-mail address: c.l.buckley@sussex.ac.uk (C.L. Buckley).
1 Joint first author.

http://dx.doi.org/10.1016/j.jmp.2017.09.004
0022-2496/© 2017 Published by Elsevier Inc.

1. Introduction

The brain sciences have long searched for a 'unified brain theory' capable of integrating experimental data relating to, and disclosing the relationships among action, perception, and learning. One promising candidate theory that has emerged over recent years is the 'free energy principle' (FEP) (Friston, 2009, 2010c). The FEP is ambitious in scope and attempts to extend even beyond the brain sciences to account for adaptive biological processes spanning an enormous range of time scales, from millisecond neuronal dynamics to the tens of millions of years span covered by evolutionary theory (Friston, 2010b, c).

The FEP has an extensive historical pedigree. Some see its origins starting with Helmholtz' proposal that perceptions are extracted from sensory data by probabilistic modelling of their causes (Von Helmholtz & Southall, 2005). Helmholtz also originated the notion of thermodynamic free energy, providing a second key inspiration for the FEP.[2] These ideas have reached recent prominence in the 'Bayesian brain' and 'predictive coding' models, according to which perceptions are the results of Bayesian inversion of a causal model, and causal models are updated by processing of sensory signals according to Bayes' rule (Bubic, von Cramon, & Schubotz, 2010; Clark, 2013; Knill & Pouget, 2004b; Rao & Ballard, 1999). However, the FEP naturally accommodates and describes both action and perception within the same framework (Friston, Daunizeau, Kilner, & Kiebel, 2010), thus others see its origins in 20th-century cybernetic principles of homoeostasis and predictive control (Seth, 2015).

A recognisable precursor to the FEP as applied to brain operation was developed by Hinton and colleagues, who showed that a function resembling free energy could be used to implement a variation of the expectation–maximisation algorithm (Neal & Hinton, 1998), as well as for training autoencoders (Hinton & Zemel, 1994a) and learning neural population codes (Zemel & Hinton, 1995). Because these algorithms integrated Bayesian ideas with a notion of free energy, Hinton named them as 'Helmholtz machines' (Dayan, Hinton, Neal, & Zemel, 1995). The FEP builds on these insights to provide a global unified theory of cognition. Essentially, the FEP generalises these results by noting that all (viable) biological organisms resist a tendency to disorder as shown by their homoeostatic properties (or, more generally, their autopoietic properties), and must therefore minimise the occurrence of events which are atypical ('surprising') in their habitable environment. For example, successful fish typically find themselves surrounded by water, and very atypically find themselves out of water, since being out of water for an extended time will lead to a breakdown of homoeostatic (autopoietic) relations. Because the distribution of 'surprising' events is in general unknown and unknowable, organisms must instead minimise a tractable proxy, which according to the FEP turns out to be 'free energy'. Free energy in this context is an information-theoretic construct that (i) provides an upper bound on the extent to which sensory data is atypical ('surprising') and (ii) can be evaluated by an organism, because it depends eventually only on sensory input and an internal model of the environmental causes of sensory input. While at its most general this theory can arguably be applied to all life-processes (Friston, 2013), it provides a particularly appealing account of brain function. Specifically it describes how neuronal processes could implement free energy minimisation either by changing sensory input via action on the world, or by updating internal models via perception, with implications for understanding the dynamics of, and interactions among action, perception, and learning. These arguments have been developed in a series of papers which have appeared over the course of the last several years (Adams, Shipp, & Friston, 2013; Carhart-Harris & Friston, 2010; Friston, 2005, 2008a; Friston, Daunizeau, & Kiebel, 2009; Friston et al., 2010; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Friston & Kiebel, 2009a, b; Friston, Kilner, & Harrison, 2006; Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007; Friston & Stephan, 2007; Friston, Stephan, Li, & Daunizeau, 2010; Friston, Trujillo-Barreto, & Daunizeau, 2008; Pezzulo, Rigoli, & Friston, 2015).

The FEP deserves close examination because of the claims made for its explanatory power. It has been suggested that the FEP discloses novel and straightforward relationships among fundamental psychological concepts such as memory, attention, value, reinforcement, and salience (Friston, 2009). Even more generally, the FEP is claimed to provide a "mathematical specification of 'what' the brain is doing" (Friston, 2009 p.300), to unify perception and action (Friston et al., 2010), and to provide a basis for integrating several general brain theories including the Bayesian brain hypothesis, neural Darwinism, Hebbian cell assembly theory, and optimal control and game theory (Friston, 2010c). The FEP has even been suggested to underlie Freudian constructs in psychoanalysis (Carhart-Harris & Friston, 2010).

Our purpose here is first to supply a mathematical appraisal of the FEP, which we hope will facilitate evaluation of claims such as those listed above; note that we do *not* attempt to resolve any such claims here. A mathematical appraisal is worthwhile because the FEP combines advanced concepts from several fields, particularly statistical physics, probability theory, machine learning, and theoretical neuroscience. The mathematics involved is non-trivial and has been presented over different stages of evolution and using varying notations. Here we first provide a complete technical account of the FEP, based on a history of publications through which the framework has been developed. Second we provide a complete description of a simple agent-based model working under this formulation. While we note that several other agent based models have been presented, e.g see (Friston et al., 2010), they have often made use of existing toolboxes which, while powerful, have perhaps clouded a fuller understanding of the FEP. Lastly we use our account to identify the assumption structure of the FEP, highlighting several instances in which non-obvious assumptions are required.

In the next section we provide a brief overview of the FEP followed by a detailed guide to the technical content covered in the rest of the paper.

[2] Thermodynamic free energy describes the macroscopic properties of nature, typically in thermal equilibrium where it takes minimum values, in terms of a few tractable variables.

Please cite this article in press as: Buckley, C.L., et al., The free energy principle for action and perception: A mathematical review. Journal of Mathematical Psychology (2017), http://dx.doi.org/10.1016/j.jmp.2017.09.004.