Studies in History and Philosophy of Science 41 (2010) 353–362

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

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

Inference to the hypothesis of extended cognition

Mark Sprevak

Department of Philosophy, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK

ARTICLE INFO

Keywords:
Extended mind
Extended cognition
Embedded cognition
Externalism
Inference to the best explanation
Functionalism

ABSTRACT

This paper examines the justification for the hypothesis of extended cognition (HEC). HEC claims that human cognitive processes can, and often do, extend outside our head to include objects in the environment. HEC has been justified by inference to the best explanation (IBE). Both advocates and critics of HEC claim that we can infer the truth value of HEC based on whether HEC makes a positive or negative explanatory contribution to cognitive science. I argue that IBE cannot play this epistemic role. A serious rival to HEC exists with a differing truth value, and this invalidates IBEs for both the truth and the falsity of HEC. Explanatory value to cognitive science is not a guide to the truth value of HEC.

1. Introduction

How much of your mind is inside your head? The *hypothesis of extended cognition* (HEC) claims that important aspects of one's mental life spill outside one's head into objects in the environment. It is commonly remarked that personal computers, calendars, notebooks, and to-do lists play a pervasive role in our lives. Such objects are in intimate feedback with our thought processes, and they guide our action in direct, and often undeliberated, ways. HEC claims that these intimate and action-guiding relationships result in those external objects being part of our cognitive processes. A fluently deployed laptop computer, iPhone, Filofax, or diary may be part of the substrate of one's mental life, in a similar manner as the neural resources inside one's head. External objects can, just like one's neural activity, constitute the realisation base of one's cognitive processes.

HEC appears to entail a radical refactoring of the mind as it is conceived in psychology, cognitive science, and philosophy of mind. If HEC is right, then those disciplines as traditionally pursued mistake their subject matter. The mind is not located inside the organism, but spread between the organism and environment. If one wishes to describe the present state of the organism's mind, or the evolution of the organism's mind over time, one must describe the organism plus its environment. A psychology or philosophy of mind that confined itself only to cognitive activity inside the organism would be impoverished along roughly the same lines as a psychology that confined itself to only one part of the brain.

One of the most influential strategies for arguing for HEC has been *inference to the best explanation* (IBE).[1] On this view, HEC is justified by its explanatory pay-off for cognitive science. The explanatory virtues of HEC for the practice of cognitive science argue for HEC's truth. IBE counsels to infer the hypothesis that best explains the data, provided that explanation meets some minimum standard for adequacy. According to Lipton (2004), IBE is central to inferential practice in science. Lipton distinguishes between two types of IBE: 'Inference to the Likeliest Explanation' and 'Inference to the Loveliest Explanation'. Inference to the Likeliest Explanation accurately describes our aspirations—we typically wish to infer the likeliest explanation—but by itself it is uninformative, it is not an effective epistemic strategy because it gives us no clue how to work out which hypothesis is the likeliest. A version of IBE that we are capable of acting on is Inference to the Loveliest Explanation. Inference to the Loveliest Explanation says that explanatory properties are a guide to likeliness. Inference to the Loveliest Explanation counsels to infer the hypothesis that provides the best (loveliest) explanation, where

E-mail address: mds26@cam.ac.uk

[1] Clark & Chalmers (1998); Clark (2008, 2007).

0039-3681/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2010.10.010

loveliest is understood in terms of explanatory properties like scope, simplicity, unification, fruitfulness, and mechanisation. An advocate of IBE claims that these properties, which make for a lovely explanation, are also a guide to truth.[2]

Instances of IBE are not hard to find. In *The Origin of Species*, Darwin cited a large array of facts, including the geographical distribution of species and the existence of atrophied organs, that are elegantly explained by the theory of evolution by natural selection, but poorly explained, or not explained at all, by rival hypotheses. In the sixth edition, Darwin wrote:

> It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified. It has recently been objected that this is an unsafe method of arguing; but it is a method used in judging of the common events of life, and has often been used by the greatest natural philosophers. (Darwin, 1962, p. 476)

Similarly, Lavoisier argued that we should posit a new chemical principle, oxygen, because of the explanatory benefits it would bring:

> I have deduced all the explanations from a simple principle, that pure or vital air is composed of a principle particular to it, which forms its base, and which I have named the *oxygen principle*, combined with the matter of fire and heat. Once this principle was admitted, the main difficulties of chemistry appeared to dissipate and vanish, and all the phenomena were explained with an astonishing simplicity. (Lavoiser, 1862, p. 623)

And Fresnel argued that the wave theory of light should be preferred to its rival, Newton's particle theory, because the wave theory better explains reflection, refraction, and diffraction:

> Thus reflection, refraction, all the cases of diffraction, colored rings in oblique incidences as in perpendicular incidences, the remarkable agreement between the thicknesses of air and of water which produce the same rings; all these phenomena, which require so many particular hypotheses in Newton's system, are reunited and explained by the theory of vibrations and influences of rays on each other. (Fresnel, 1866, p. 36)[3]

IBE appears to have played an important epistemic role in some of our most prized scientific inferences. IBE may not be the only way in which a scientific hypothesis is supported, but it does appear to have the ability to confer significant epistemic warrant.

Supporters of HEC argue that the best explanation of the evidence in cognitive science is the truth of HEC. HEC provides the most unified, fruitful, and elegant explanation of the empirical data. Hence, we should infer that HEC is true. Opponents of HEC employ IBE to argue for HEC's falsity.[4] They argue that HEC contributes negative explanatory value to cognitive science, and hence we should infer that HEC is false, since its falsity would better explain the data than its truth. Both critics and advocates of HEC agree that HEC's explanatory value is a guide to its truth value. They disagree about the direction in which the explanatory guide points: whether HEC's explanatory contribution to cognitive science is positive or negative. If positive, we should infer HEC's truth; if negative, we should infer HEC's falsity.

In this paper, I argue that both critics and advocates of HEC are mistaken. IBE fails as a way both of arguing for HEC, and as a way of criticising HEC. The reason is a common source of failure with IBE: the existence of a hypothesis that is a serious explanatory rival with a differing truth value. IBE is highly sensitive to the competitive context. Introducing the right kind of rival can dramatically alter the result of explanatory competitions and invalidate plausible IBEs, even if the empirical 'evidence' has not changed.[5] A reasonable IBE in one context may be rendered invalid if a better, or an equally good, rival explanation is introduced. I argue that once the right rival to HEC is considered, one can see that IBE cannot do the epistemic work that it has been claimed to do. Advocates and critics of HEC have won an unjustified sheen of plausibility for their arguments by shielding them from appropriate rivals. Once these rivals are introduced, HEC is simply not sensitive to the empirical practice of cognitive science in a way that an IBE based on that practice can bring to bear. HEC should be criticised or supported in other ways.

The argument of this paper, although primarily about the status of HEC, should interest a wider constituency than just fans and critics of extended cognition. The argument is an illustration of pitfalls with an increasingly common naturalistic move in philosophy of mind. In some quarters, there is a tendency to appeal to scientific practice as the ultimate arbiter of hypotheses. This extreme, knee-jerk, naturalism mistakenly treats contingent features of scientific practice with a higher degree of reverence than they deserve, or than scientists themselves would accord. One cannot read off metaphysics from science, or reduce metaphysical questions to questions of scientific practice. The argument below aims to show that the judgements of science (even a future cognitive science) are, by themselves, too limited to decide the extent of our mental life.

Lipton distinguishes between *descriptive* and *normative* questions concerning IBE. The descriptive question is: does IBE provide an accurate description of the actual inferential practices in science? The normative question is: does IBE propose an inferential method that is likely to take us to the truth? This paper primarily concerns the normative dimension of IBE. The question is whether explanatory value to cognitive science gives us a reason to infer the truth/falsity of HEC. One way of approaching this question would be to ask whether IBE is *generally* a reliable form of inference. In this paper, I wish to grant for the sake of argument IBE the status of being generally trustworthy. I wish to see if on the most sympathetic understanding of IBE it supports the arguments for and against HEC. Moreover, although the focus of this paper is the normative project, it also contributes to the descriptive project by highlighting specific properties accorded explanatory significance in cognitive science.

2. HEC

A tempting picture of the mind is of an entity that could, in principle, be divorced from the world and yet remain largely untouched. Descartes explored consequences of this picture when he considered the possibility that the world might be radically different, while one's mind remains the same. On such a view, one's mind *causally interacts* with the environment (via reliable or unreliable channels), but it is *constituted* largely independently of that environment. The mind could, in principle, be transplanted without significant loss into an impoverished environment. This picture has been undermined from a number of directions.

First, it faces the challenge of *content externalism*. Putnam (1975) and Burge (1979, 1986) argue that the content of certain beliefs and other mental states depends on distal features of one's environment and one's history. An exact physical duplicate in

[2] Lipton (2004), pp. 59–62, 122.

[3] The quotations are taken from Thagard (1978) using his translation.

[4] Adams & Aizawa (2007); Aizawa (2007); Rupert (2004, 2009a,b).

[5] Cf. van Fraassen (1980)'s strategy for defanging the miracle argument for scientific realism by introducing a rival Darwinian explanation of the success of scientific theories.