Published in *Studies in History and Philosophy of Science* (2010) 41: 353–362.
doi:10.1016/j.shpsa.2010.10.010

mark.sprevak@ed.ac.uk

Inference to the hypothesis of extended cognition

Mark Sprevak
University of Edinburgh

11 November 2010

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 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

1. Clark (2007, 2008); Clark and Chalmers (1998).
2. Lipton (2004), pp. 59–62, 122.

> 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

3. The quotations are taken from Thagard (1978) using his translation.
4. Adams and Aizawa (2007); Aizawa (2007); Rupert (2004, 2009a, 2009b).

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

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.

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 another environment may have different mental contents solely in virtue of its different surroundings. A second challenge is the possibility of *factive mental states*. Williamson (2000) argues that knowledge is one among many factive mental states. A factive state has true contents, and consequently factive mental states (knowing, perceiving, remembering, etc.) can depend both on how things are inside our head, and on how things are in the environment. Changes in one's environment can change one's factive mental states (e.g. transform knowledge to mere belief) without any concomitant internal changes.

HEC challenges the internalist picture from another direction. According to HEC, the environment plays an intimate and pervasive role in constituting the *mechanisms* of cognitive processing. An important part of the internalist picture is the assumption that one's mental states and processes somehow take place in one's neural tissue or body. There are a variety of ways in which this internalist thought can be given content. An identity theorist may claim that mental states are *identical* to brain states and hence, in a robust sense, are located inside the head. A functionalist may claim that mental states are *realised* by brain states, and so are similarly associated with an internal location. A property dualist may claim that mental properties are *instantiated* in the brain and nervous system, or in the whole human organism, and hence although non-physical, closely tied to an internal location. Even a substance

dualist like Descartes can make sense of the thought that cognition is an internal process: the location at which mental states enter the physical world—the pineal gland—is robustly inside the head. For the internalist, if the mechanisms of mental processes lie anywhere in space at all, they lie within the skin. HEC claims that this picture is wrong. Environmental processes support our mental activity in exactly the same way as internal activity. Our mental states are constituted in, or realised by, or instantiated by, environmental processes just as they are constituted in, or realised by, or instantiated by, neural processes. Environmental activity forms part of the mechanisms of one's mind in the same way as one's neural activity.

In contrast to content externalism, HEC makes a claim about the location of the vehicles, not the content, of mental states. Putnam's and Burge's externalism assumed that the vehicles of mental states were neural even if the factors that determined their content were external. In contrast to externalism about factive mental states, HEC can apply to any aspect of one's mental life, not just factive states like knowing or perceiving. HEC is also distinctively active in a way that the other two forms of externalism are not. Environmental features generally play an action-guiding role in HEC. Changes in the environment can trigger changes in the way in which thoughts are processed, which can cause changes in the behaviour of the cognitive agent. As far as the other forms of externalism are concerned, although the environment can affect the content or factive status of a mental state, the mental state's action-guiding role typically remains untouched.[6]

In order to state HEC, a trivial and a non-trivial way in which the mind could depend on the environment should be distinguished. The trivial form of dependence is *causal dependence*. The human brain causally depends on an endocrine system, blood supply, and being in an oxygen-rich environment. This does not, by itself, make the endocrine system, blood supply, or environmental oxygen part of the brain. Similarly, our mental states causally depend on the environment for their input: we would not enjoy the rich mental life we do if it were not for the environmental sensory input. However, this kind of dependence is compatible with mental activity being wholly internal. An internalist may admit that our mental states causally depend on the environment, but claim that the environment is needed only insofar as it provides input that brings about the right internal states, which are the true locus of mental activity. On such a view, the environment plays only an instrumental role in causing cognitive states to occur.[7]

HEC makes a stronger claim: the dependence of the mind on the environment is not just causal, it is also *constitutive*. The mechanisms of one's mind are partially made up from the environment, in a similar way as the mechanisms of one's mind are

6. See Clark and Chalmers (1998), p. 9; Clark (2009); Hurley (2010).
7. See Hurley (2010).

partially made up from activity in one's auditory cortex. Rather than being merely instrumental non-mental input, environmental states are part of the cognitive mechanism. In functionalist terms, cognitive mechanisms are *jointly realised* in both environmental activity and neural activity without any difference in kind, cognitive or mental, between the environmental and neural parts. Environmental activity is more than a convenient way of bringing about the right internal states, it is part of the mechanism of certain cognitive processes, and a necessary precondition for certain cognitive processes to exist at all.

Unsurprisingly, justifying whether something is a cause or a constitutive part of a mental process is far from easy. Initially, advocates of HEC appeared to argue for a constitutive relationship with the environment based on 'close-coupling' arguments. They pointed out the existence of tight causal loops criss-crossing between the mind and certain features in the environment, and claimed that because of the tightness of these loops, we should infer that those environmental features are part of the mind.[8] Adams and Aizawa (2001, 2007) objected that this form of reasoning commits what they call the 'coupling–constitution fallacy'. Just because x and y are tightly causally coupled, that does not show that x is part of y, or that y is part of x. A bimetallic strip inside a thermostat is tightly causally coupled to the motion of atoms in the surrounding air, but that does not show that expansion of the strip is a process that extends into the atoms in the air. Adams and Aizawa argue that the limits of the causal–constitutive boundary should instead be decided by what they call a *mark of the cognitive*: a general theory that specifies what makes something a cognitive process. Adams & Aizawa proposed two necessary conditions as part of such a theory: a cognitive process should involve non-derived content, and it should have a fine-grained causal structure similar to that of actual human cognitive processes. Both criteria have been criticised.[9] There is currently no consensus on what should count as a mark of the cognitive.

The result has been a stalemate concerning whether coupling arguments for HEC are successful or not. No proposed mark of the cognitive has proved less controversial

8. Attributing this argument to Clark and Chalmers (1998) has become common, but it not unproblematic because the causal/constitutive distinction was not explicitly drawn at the time. The argument is attributed based on their discussion of coupling in §§3,4. See Adams and Aizawa (2007), pp. 89–90 for discussion.

9. See Clark (2005, 2010) on the non-derived content condition and Clark (2010); Sprevak (2009) on the causal structure condition. Rupert (2009a, 2009b) proposes another mark of the cognitive—cognitive processes are persisting sets of integrated packages—also controversial, but discussion of which would take us too far afield.

than the cases that it is supposed to decide.[10] In order to break this stalemate, there has been a recent turn to scientific practice. Despite our lack of agreement on a theoretical mark of the cognitive, psychology makes decisions all the time about whether a causal process is an input to a cognitive process or a part of the cognitive process. We appear to have an implicit grasp of the causal–constitutive split from the typings of cognitive processes that are arrived at in psychology. These typings are governed in large part by questions of explanatory value: whether a process is typed as two processes or a single psychological process depends on what best serves the interests of psychological explanations. This explanatory practice promises to provide a way of demarcating the causal–constitutive split. Hurley nicely summarises the strategy:

> Criteria of the mental vary widely (if not wildly) across different theorists; it isn't even clear what agreed work such criteria should do. Yet psychology continues on its way with a rough and ready sense of what it wants to explain, generating good explanations. The issues between internalism and externalism should resolved bottom up by such scientific practice, not by advance metaphysics: by seeing whether any good psychological explanations are externalist, not by deciding on a criterion of the mental and using it to sort explanations as constitutive or not. In this context, I'm aware of no appropriate criterion independent of good explanations. (Hurley 2010, pp. 106–107)

In a similar vein, Clark (2007) argues that some coupling relations (self-stimulating loops), but not others, serve as single explanatory units in psychology, and thereby have a constitutive rather than merely causal role (pp. 184–185). Adams and Aizawa (2007) and Rupert (2009a, 2009b), in addition to proposing theoretical marks of the cognitive, argue that those marks should be justified by the explanatory practice of psychology.[11] Chalmers, after considering a purely causal gloss of external factors, says: 'Ultimately, the proof is in the pudding. The deepest support for [externalism] comes from the explanatory insights that the extended mind perspective yields' (Chalmers 2008, p. 16).

10. Notably, stalemate does not mean that internalism wins by default. To assume this would be to commit what Hurley (2010) calls the 'causal–constitutive error error': claiming that HEC gives an unjustified constitutive role to external factors, while assuming without independent justification that the constitutive boundary falls inside the skin. If neither camp can draw the causal/constitutive boundary, why assume that it favours internalism?

11. 'In the end empirical research should decide this question: we should commit resources to the framework of extended cognitive systems, apply the extended view in the study and the lab, and see whether doing so generates a flourishing research program in cognitive science' (Rupert 2004, p. 425).

The close-coupling argument initially appeared to be distinct from the direct IBEs for HEC described below, now it appears to rest on explanatory concerns. Therefore, in recent years IBE has become increasingly important in debates about HEC. IBE arguments have to do double duty. First, argue directly for the truth of HEC based on its explanatory pay-off (as described below). Second, settle whether the coupling relations involved are constitutive or merely causal. IBE has taken centre stage as one of the key arguments for and against HEC. It is to these IBE arguments that we now turn.

3 Finding HEC's explanatory rivals

Inference to the best explanation counsels to infer the best among all the explanations we can generate of the data (provided that best explanation is good enough). The competition between explanations should be free and fair: we should not bias the field by deliberately ignoring certain known alternatives. In other words, for an IBE to be valid it is not enough that a hypothesis beat *some* competitors, it should beat all known competitors. For if a better explanation were in hand, it would have a stronger claim to be inferred.

There is no guarantee that we have the best explanation in hand. At a given moment, we can only think of a limited number of explanations. If IBE is to be an effective epistemic strategy, it should, at most, counsel us to infer the best explanation we can *currently generate*; it cannot ask us to do the impossible. This raises the possibility that tomorrow we may think of a new explanation that upsets a previous explanatory competition. Consequently, inferences to the best explanation are defeasible, in at least two ways. First, new data coming in can affect the success or failure of existing explanations. An existing explanation may become more or less plausible in the light of new data. Second, creating a new explanation can defeat a previous IBE by usurping the title of best explanation. A plausible IBE can be rendered invalid by the introduction of a new competitor, even in the absence of any new data. In the next two sections, I argue that creating the right competing hypothesis defeats IBEs both for and against HEC.

What are the competing explanations to HEC?

One competitor to HEC is a version of internalism that I call the *hypothesis of internal cognition* (HINT). According to HINT, psychological processes do not extend outside the head, and they can be explained and studied largely in isolation from their environment. The input and output of a psychological process may be located in the environment, but the specification and explanation of psychological mechanisms is largely a story about internal activity. Just as the explanation of

the mechanisms of a personal computer is typically given irrespective of the environment in which the computer is located (save for specification of its inputs and outputs), so the explanation of the mechanisms of human psychology should be given largely irrespective of the environment in which the organism happens to be located. This is not to say that under HINT no explanation of psychological mechanisms would appeal to the environment. The environment will undoubtedly play a role in answering *why* we have certain cognitive processes (e.g. as part of an evolutionary story, or a developmental story). But the specification and explanation of *what those psychological processes are* is nevertheless a story about internal mechanisms.[12]

HINT's influences run deep and are evident in psychological explanation at least as far back as Descartes. Post-war psychology has been particularly heavily influenced by HINT, not least because it fits naturally with the model of the mind as a computer.[13] HINT enjoys obvious appeal: we appear to carry our cognitive mechanisms around with us, and we appear to be capable of deploying our cognitive mechanisms across different environments. Surely then our cognitive mechanisms are internal to us, and their study a matter of specifying internal processes? HINT has consequences for psychological practice: it makes cognitive science particularly apt for laboratory work and analyses that abstract cognitive mechanisms from their natural environment.[14]

HEC and HINT are rivals, but HINT is not the only rival to HEC. Another is introduced by Rupert (2004): the *hypothesis of embedded cognition* (HEMC).[15] Like HEC, HEMC claims that cognition depends intimately, and often in unexpected ways, on external props and the structure of the environment. Like HEC, HEMC claims that an adequate specification and explanation of those cognitive mechanisms should be a complex story involving both internal activity and environmental activity. HEMC agrees with HEC that the environment is not just a net input or output of a cognitive process, it can be an essential part of the intermediate processing. The one point on which HEMC differs from HEC is that HEMC does not claim that the relevant environmental features are thereby mental or cognitive. HEMC agrees with HEC about the respects in which the environment enters into psychological explanation; but it departs from HEC in that it does not take on HEC's claim that those environmental features are also mental.

12. Cf. Marr (1982)'s distinction between (i) the 'ecological' question of *why* we have certain cognitive processes (answered at the 'computational level'), and (ii) *how* those cognitive processes work (answered by a mechanistic story at the 'algorithmic level'), and for which, unlike (i), an internalist story is told.

13. Cf. Fodor (1980)'s methodological solipsism.

14. See Chirimuuta and Gold (2009).

15. Rupert develops HEMC from McClamrock (1995). I take liberties with developing Rupert's version of HEMC.

HEMC splits two claims that are combined in statements of HEC and HINT: a claim about cognitive *mechanisms* and a claim about the *mechanistic explanation* of cognition. HEMC does not claim that the mechanisms of cognition are external; it only claims that the mechanistic explanation of how psychological processes work is not a purely internal story. Such a split need not be peculiar to cognition. For instance, one might think that the mechanisms of the human liver are internal to the liver, but that those mechanisms are so bound up with other bodily systems that the story about how they work cannot be served by a purely internal narrative. Similarly, an advocate of HEMC may claim that cognitive mechanisms are internal, but that the mechanistic explanation of how they work is a complex story involving both internal activity and environmental resources. According to HEMC, one's cognitive processes stop at the skull, but a description of what takes place inside the skull is not sufficient as an explanation of how one's cognitive mechanisms work.

HEMC has similar consequences to HEC for the way in which psychology should be practised. HEMC colours one's view of what goes on inside the thinking subject and opens up the possibility that the subject 'off-loads' some of the work involved in cognition onto environmental resources, so that she can get away with simpler internal mental processes, or internal processes of a different character. The study of psychological mechanisms according to HEMC should not just be a study of internal activity but include how we exploit features of the environment. In contrast, HINT requires that nearly all the work of cognition be done in the head. A HEMC-based psychologist can explain how a cognitive process works by showing how interaction with the environment is harnessed as part of the cognitive processing, it does not—as HINT does—adopt a picture where only internal processing is relevant to the explanation of the mechanisms of cognition.

HEC, HEMC, and HINT are three competing hypotheses. HEMC claims that extra-cranial features play an essential role in the specification and explanation of cognitive mechanisms. HEC claims that those same features play that same role, *and* they are also mental. HINT claims that the extra-cranial features are neither mental nor play an essential role in explaining how cognitive processes work (save for functioning as inputs or outputs to the mechanism, or as general background conditions for the internal mechanisms to function at all).

It is worth distinguishing between a stronger and a weaker version of HEMC. Strictly speaking, the version of HEMC described above (and given by Rupert (2004)) is compatible with HEC. This version of HEMC does not deny that cognition extends, it just does not, unlike HEC, affirm it. This raises a potential puzzle for IBE arguments. IBE is a mechanism for selecting among rivals, but HEC and HEMC are logically compatible, and so are not genuine rivals. A way around this difficulty, the way adopted by Rupert and others, is to combine HEMC with a form of internalism, and

so have an overall package of views claiming that cognitive mechanisms are internal. This seems the best way of interpreting HEMC: the purpose of HEMC is to gain the benefits of HEC, but keep cognition internal. Alternative interpretations are unpromising: combining HEMC with HEC does not result in a new position, and combining HEMC with agnosticism about internalism/externalism does not yield an IBE candidate capable of showing that HEC is true or false. Therefore, at least for the purposes of IBE arguments, I will construe HEMC as a variety of internalism.

4 Explanatory value arguments for HEC

Clark and Chalmers (1998) give the following IBE for HEC:

> By embracing an active externalism [HEC], we allow a more natural explanation of all sorts of actions. One can explain my choice of words in Scrabble, for example, as the outcome of an extended cognitive process involving the rearrangement of tiles on my tray. Of course, one could always try to explain my action in terms of internal processes and a long series of 'inputs' and 'actions', but this explanation would be needlessly complex. (pp. 9–10)

In this case, Clark and Chalmers claim that a psychological explanation should not just focus on the internal activity that gives rise to behaviour, but also on the ways in which cognitive activity relies on the structure of the environment to achieve the cognitive agent's ends. The way in which I rearrange my Scrabble tiles is not just a behavioural output, but a strategy on my part of structuring the world to help me think. The intended contrast is between explaining mental activity in purely internal terms, and giving a mixed story in terms of a mechanism that includes both internal activity and environmental props.

If successful, this IBE may target HINT, but it does not discriminate between HEC and HEMC. Both HEMC and HEC acknowledge that an internal story is insufficient to explain all psychological processes, and both hypotheses acknowledge that cognitive processes rely on the structure of the environment. Clark and Chalmers' 'active externalism' mode of explanation is available in both cases. Furthermore, it is far from clear what additional explanatory benefit is gained by HEC's claim that the Scrabble tiles are *mental* or *cognitive*, rather than saying, as HEMC does, that they are non-mental props essential to the explanation of successful cognition. Therefore, the IBE above may select between HINT and HEC or HEMC, but it does not select between HEC and HEMC, and so does not tell distinctively in favour of HEC.

Another argument from Clark and Chalmers is more promising. I call this the *transcranial kinds* argument. Clark and Chalmers present two cases of supposed belief. The first involves Inga, a normal human subject who hears of an exhibition at the Museum of Modern Art (MoMA). Inga thinks, recalls that MoMA is on 53rd St., and sets off. Otto, who suffers from a mild form of Alzheimer's and always writes down useful information in his notebook, also hears of the exhibition. Otto retrieves the address from his notebook, and sets off.

Clark and Chalmers claim that these two cases fall under a common psychological explanation. Just as Inga arrived at 53rd St. because she wanted to go to MoMA and believed that it was there, so Otto arrived at 53rd St. because he wanted to go to MoMA and believed that it was there. The state that functioned as Otto's belief was not entirely inside his head—it was spread between his neural activity and his notebook—but that state still combined with Otto's desires and drove his action in the same way as Inga's internal belief.

According to Clark and Chalmers, treating Otto's extended state as a belief unifies the psychological explanation of Otto and Inga in a valuable way. It allows one to see a common psychological action at work, irrespective of whether the agent relies only on internal resources, or off-loads work onto the environment. Moreover, the alternative—explaining Otto's success in terms of his internal beliefs about the notebook—seems needlessly complex. The notebook is a constant in Otto's life, just like Inga's internal memory. It appears redundant to point to Otto's notebook in every explanation of his action. A common pattern of belief–desire explanation beckons in both cases. Positing extended beliefs provides the best—most unified and most elegant—explanation of the agents' behaviour:[16]

> By using the 'belief' notion in a wider way, it picks out something more akin to a natural kind. The notion becomes deeper and more unified, and it is more useful in explanation.
>
> (Clark and Chalmers 1998, p. 14)

Clark and Chalmers claim that by introducing psychological kinds that cross-cut the organism's boundary—*transcranial cognitive kinds*—cognitive science becomes more powerful and unified. The explanatory value of transcranial cognitive kinds is given as an argument for their existence.

One way of objecting to this IBE argument for HEC is to argue that transcranial kinds do *not* bring such reputed explanatory benefits to cognitive science. This is the route pursued by Rupert (2004).[17] Rupert argues that the valuable generalisations

16. Clark and Chalmers (1998), pp. 12–14.
17. See for example, pp. 420–421.

in cognitive science concern the *particular methods* by which individual subjects access and process information. A psychology that treats internal and transcranial cases as of the same psychological kind will, by necessity, be uninformative about the details of those individual methods. In order for Otto and Inga to instantiate the same psychological kind, the fine-grained details of their individual memory mechanisms have to be glossed over. But according to Rupert, those fine-grained details are precisely what cognitive science should uncover. Consequently, adding transcranial kinds obstructs progress in cognitive science rather than aids it.

As Rupert admits, the problem with this objection is that an advocate of HEC is not committed to *dispensing* with individualistic or fine-grained cognitive kinds, only to *adding* transcranial kinds to the explanatory resources of cognitive science. It is no part of HEC to outlaw traditional psychology or ignore the fine-grained ways in which Otto and Inga differ. Rather, HEC's claim is that traditional individualistic psychology should be seen as part of a wider story about the recruitment of resources that may or may not be internal. An advocate of HEC can agree with Rupert that Otto and Inga differ in their fine-grained memory details, and that in certain contexts such differences matter a great deal to psychological explanation. Her claim is that *in addition to those fine-grained differences*, Otto and Inga share an overarching mental kind.

Another difficulty with a flat-footed denial that transcranial kinds have explanatory value is that transcranial kinds fall under a wider explanatory strategy in psychology. Psychologists often 'black-box' a cognitive mechanism and abstract away its details. Black-boxing is established as a useful explanatory manoeuvre. It allows one to see a common architecture at work, and common mechanisms shared between different sub-systems. It has become a primary strategy in understanding how human psychology works. If ignoring fine-grained details, black-boxing, is valid in internal cases, why not in extended cases too? Black-boxing promises to capture an account of the action involved when humans rely on, and switch between, internal and external memory resources. This does not rule out a supplementary account of the differences between those memory resources. But it adds the ability to understand a higher-level pattern that they share. Humans appear to switch between information stored internally and information off-loaded onto environmental resources. Transcranial kinds allow one to understand how there could be a common architecture at work governing this process.[18]

I argue that the problem with transcranial kinds is not, as Rupert contends, that they are explanatorily worthless, but that whatever explanatory value transcranial kinds bring to cognitive science, they do not select between HEC and HEMC.

18. See Gray and Fu (2004); Gray et al. (2006), Ballard, Hayhoe and Pelz (1995); Ballard et al. (1997), and the discussion below for examples of this kind of work in psychology.

Clark and Chalmers' transcranial kinds argument for HEC relies on the assumption that transcranial kinds are mental kinds. But what is the explanatory value in believing this? Suppose that transcranial kinds are *hybrid kinds*: part mental, part non-mental. Hybrid kinds are compatible with HEMC. Hybrid kinds also appear equipped to take on the explanatory duties described above. The explanatory benefits that Clark and Chalmers describe for transcranial kinds stem from their transcranial nature, not from their supposedly mental nature. Indeed, it is far from clear that transcranial kinds being *mental* or *cognitive* does any explanatory work at all.

Clark and Chalmers claim that transcranial kinds unify the psychological explanation of Otto and Inga. But this unified form of explanation is available under either the hybrid kind or pure mental kind hypotheses. In both cases, one can assert that Otto and Inga share a psychologically-significant state. Why should this state be 100% mental in both cases? In order to achieve unified explanation, it is sufficient they share *some* psychologically-significant state, and hence fall under same psychological explanation. It is a further question whether what they share is mental or a mix of mental and non-mental elements in each case.

Notably, being *psychologically-significant* is not sufficient to make a state mental. Psychology contains a mix of mental and non-mental elements significant for explanation: for example, the causes and effects of mental states are often not mental themselves, and yet are still significant in psychological explanations. If transcranial kinds are to be included in the taxonomy of cognitive psychology, why assume that they fall cleanly under the heading of pure mental states?

Inga's psychological state is not identical in all respects to Otto's psychological state. In order to satisfy the same psychological explanation they must share some features in common. But why assume that their common properties include a shared *mental* nature, rather than, say, a shared role in driving action under a relatively abstract description? The explanatory value to cognitive science comes from positing a transcranial state that has certain psychologically-significant features, not from assuming that such a state is wholly mental.[19]

Transcranial kinds allow one to illustrate a common psychological architecture across differences in resources, internal or external. But a description of a common

19. One might object that *because* of a transcranial state's psychologically-significant commonalities with internal states (e.g. it is action guiding, combines with internal desires in the right way, etc.), the transcranial state *ipso facto* counts as mental. This is, in abbreviated form, the functionalist argument for HEC: since the transcranial state satisfies some minimal functional specification to be a mental state, it should count as a mental state. I argue at length against the functionalist argument, and the closely-related parity principle, in Sprevak (2009). My concern here is to address whether *IBE arguments* lend any special, independent, support to HEC. My claim is that they do not. The virtues canvassed by Clark and Chalmers for HEC (unification, etc.) do not decide between HEC and HEMC since they are equally enjoyed by both hypotheses.

architecture can be given without settling whether the resource abstracted over is mental in both cases. In some instances, a recall mechanism might take advantage of mental resources, in others it might use a non-mental resource. The explanatory value of transcranial kinds comes from allowing us to generalise about how mental processes involve, and switch between, internal and external resources. It does not derive from the assumption that the resource involved in each case is mental.

Transcranial kinds allow us to explain Otto's success without repeatedly mentioning his beliefs about his notebook. But again this leaves open whether the state responsible for Otto's success is mental. All that is required is that Otto and Inga share some psychologically-significant features responsible for their successful action. As indicated above, this is available on either hypothesis.

A final potential explanatory benefit of transcranial kinds is that they allow one to classify traditional sensory input, internal state, behavioural output cycles as single explanatory units in psychology. But treating entire sensorimotor loops as single explanatory units is compatible with holding that the input and output sections of the loops are non-mental, albeit psychologically significant. An advocate of HEMC can obtain whatever explanatory benefits are to be gained by appealing to whole loops as single explanatory units, without being committed to the sensorimotor loops being 100% mental in all their parts.

In short, the explanatory work done by transcranial kinds comes from their transcranial nature. No additional bonus comes from the claim that the transcranial kinds are mental. What is the explanatory edge in believing that Otto's notebook is *mental*, over and above believing it to be a psychologically-significant feature for guiding his action? To achieve the benefits indicated by Clark and Chalmers, psychology only needs to be reformed to include transcranial kinds. The additional assumption that transcranial kinds be mental kinds is otiose.

One might object that hybrid kinds are nevertheless undesirable, or at least less desirable, than pure kinds. Hybrid kinds cross the boundary between natural kinds, in this case, the mental/non-mental boundary. One may say that all else being equal, it is preferable to preserve these natural kind structures in science. For example, one might argue that in science we should prefer hypotheses that do not traffic in 'jade', but rather distinguish jadeite from nephrite. Similarly, one should prefer a hypothesis that does not deal in hybrid mental/non-mental kinds in favour of one that deals in purely mental or purely non-mental kinds. A hypothesis that groups things that are relevantly similar together, and distinguishes things that are relevantly different, has an explanatory edge.

The problem with this objection is that it asserts that such an explanatory edge exists, whereas we appear to have every reason to think that it does not. HEMC shows that we can meet our explanatory obligations, and remain neutral on the question of

mentality. HEMC appears to show that *there is no* explanatory edge in preserving the natural kind structure. If there are no explanatory benefits, it simply does not matter whether one preserves the relevant natural kind structure or not. Again, the case is like 'jade'. If 'jade' were to play no explanatory role in science, it would not matter whether we call two samples of rock 'jade' or not. Similarly, since the notion of mentality appears to play no explanatory role above, there is no benefit to be gained in using mental kinds over hybrid kinds. The natural kind facts about which HEC and HEMC differ are just not doing any explanatory work.

A more promising line of objection is to say that mental kinds are preferable to hybrid kinds because positing purely mental kinds is, in some sense, the simpler option. Appealing to mental kinds yields explanations that employ fewer, or less complex, types of entities. A uniformly mental story is simpler and more elegant than a relatively ugly mixed psychological explanation based on hybrid kinds. I will bracket this concern until Section 5, where I discuss it in the context of IBEs against HEC.

Clark (2007) has given two other IBEs for HEC, but it is hard to see how they fare any better against HEMC.

The first argument is based on *cognitive impartiality*.[20] The argument derives from recent empirical work by Gray and Ballard.[21] Their work appears to show that the brain does not 'care' whether its operations are performed internally or externally, so long as they satisfy some cost–benefit function, which may not, and often does not, privilege the skin boundary. Therefore, HEC has the virtue that it lets us see the boundaries of skin and skull as the brain sees them: as functionally transparent. HEMC obscures the view by erecting a barrier that the system itself does not care about. However, even if Clark's interpretation of the empirical work is correct, it is hard to see how this is an argument for HEC over HEMC. It is no part of HEMC to deny that the brain can off-load operations onto the external environment, or that such dependencies cannot be negotiated by a cost–benefit function.[22] HEMC is compatible with seeing the division of labour between internal activity and environmental resources as governed in the same way as HEC does, by a cost–benefit function that does not privilege the skin. The only difference is that HEMC does not claim that the extra-cranial material is mental, whereas HEC claims that it is. So it is not clear how a cost–benefit explanation of the brain's internal/external resource switching is the property of HEC but not HEMC.

20. Clark (2007), pp. 171–176, 190.

21. Gray and Fu (2004); Gray et al. (2006), Ballard, Hayhoe and Pelz (1995); Ballard et al. (1997).

22. Neither does HEMC claim that the division of labour is a conscious decision on the part of the agent, or that it involves a Cartesian Theatre. (Clark 2007, pp. 190–191)

Clark's second argument is based on *self-stimulating loops*.[23] Clark claims that certain kinds of causal relation to the environment, self-stimulating loops, are best understood as single explanatory units in psychology, and therefore as entirely cognitive. A self-stimulating loop is a process in which the agent creates output (speech, gesture, written words), which is recycled as sensory input (hearing, touch, vision), which drives the cognitive process along, producing with it more output. Clark's analogy is with a turbo-charged engine: exhaust flow from the engine is fed back into the engine in order to spin a compression pump allowing for more powerful internal combustions and producing greater power. Just as the whole turbo-charging cycle counts as part of the automobile's power-generating mechanism, so the whole self-stimulating loop should count as part of the agent's cognitive mechanism when explaining its successful action. But again, it is hard to see how this is an argument for HEC over HEMC. Viewing a self-stimulating loop as a single explanatory unit in psychology is compatible with HEMC as well as HEC. The difference between the two hypotheses is whether the loop involves exclusively mental kinds or a hybrid mix of mental and non-mental kinds. And on this question, explanatory utility does not select between HEC and HEMC.

5 Explanatory value arguments against HEC

Rupert (2004, 2009b) claims that there is a successful IBE against HEC. According to Rupert, HEMC is a *better* explanatory hypothesis to HEC. Consequently, we should infer HEMC's truth and HEC's falsity. The strategy is disconfirmation by rival support. HEMC's truth, and HEC's falsity, provides a better explanation of the empirical data in cognitive science than HEC's truth. So while Clark and Chalmers argue that explanatory concerns support HEC, Rupert argues that explanatory concerns undermine HEC.

My claim is that both sides are mistaken. The failure of an IBE for HEC should give no comfort to a critic who wishes to mount an IBE against HEC. The lesson from the discussion above is not that HEMC wins the HEC/HEMC contest, but that an IBE based on scientific practice is the wrong tool to decide between HEC and HEMC. A cognitive scientist could swap between HEC and HEMC with negligible net change in explanatory value. Neither HEC nor HEMC are clear explanatory winners, so one cannot use their explanatory value as a guide to the truth.

Rupert argues that HEMC is explanatorily better than HEC because HEMC is a *simpler* and more *conservative* hypothesis:

23. Clark (2007), pp. 183–185, p. 190.

> If the cases canvassed here are any indication, adopting HEC … at the very best, yields only an unmotivated reinterpretation of results that can, at little cost, be systematically accounted for within a more conservative framework. (Rupert 2004, p. 390)

> … all other things being equal, we should endorse HEMC over HEC, by dint of the methodological principle of conservatism. (p. 395)

> If HEMC accounts for the results that impress advocates of HEC, the more conservative, simpler HEMC wins the day. (Rupert 2009b)

Conservativeness and simplicity are distinct virtues. Conservativeness is backward looking: it speaks to a better fit of the new theory with the old picture. Simplicity is sideways looking: it speaks to the merits of the new theory with respect to other live current alternatives. Let us consider each virtue in turn.

First, conservativeness. Rupert is right that HEMC is more conservative than HEC: HEC departs from the traditional view by claiming that environmental features are mental, while HEMC preserves the traditional internalist framework. However, it is unclear how much weight should be accorded to this win, or even if it should weigh anything at all in this debate.

First, the internalist view that HEMC preserves is primarily a product, not of empirical science, but our folk conception of the mental. It would be a mistake to assume that the traditional internalist picture is the output of a mature scientific theory. It is not so much that the externalist question has been considered by science, and an internalist answer given, but that the internalist conception has been imported into cognitive science from prior folk conceptions of the mind. For the most part, the question concerning HEC has not been explicitly considered by empirical science; no existing scientific theory aims to answer the question. Internalism has been taken for granted in psychology in a largely unargued for manner. Conservativeness is a virtue when one has good reasons for believing the old theory, e.g. when the old theory is a tried and trusted scientific theory. However, when the old view is just our folk conception in scientific clothing, it is unclear that conservativeness with this view should weigh heavily at all. Indeed, there is plenty of evidence that conservativeness with the folk conception has been a systematically misleading guide to the truth in science. An argument for HEMC's truth based solely on its conservativeness offers flimsy support to HEMC indeed.

Second, and more worryingly for Rupert, appeal to conservativeness to support HEMC is tantamount to begging the question against HEC. The issue at stake is whether we should reform our folk view in the way suggested by HEC. The reply on

offer is that it fits better with the folk conception not to reform, and hence by dint of the virtue of conservativeness, we should preserve the old view. But the advocate of HEC already freely admits that her view departs from folk conception—indeed, that is the point of her view. The question is whether her new conception is true. Objecting that one thinks that she is wrong because she is proposing reform, and one disfavours reform, is to fail to take her claim seriously. One wants to know whether the proposed reform is true. Being told that it is reform is not news. In short, it is already granted by both sides that HEC is less conservative then HEMC. Some other explanatory virtue must come into play. Therefore, an IBE against HEC cannot be an IBE based on conservativeness.

Simplicity seems a more promising line of attack. Rupert claims that HEMC is more parsimonious than HEC, and consequently that HEMC is the simpler hypothesis. Both HEC and HEMC posit the same causal processes and interactions between the brain and external environment that are relevant to the explanation of cognition. However, in addition HEC claims that some of these causal processes are extended *mental* or *cognitive* processes. Yet HEC appears to reap no explanatory rewards from this addition; as we saw above, the same explanatory work can be done without any ontological inflation. If the extra content of HEC—appeal to specifically *mental* or *cognitive* environmental entities—does no extra work, we should prefer the more parsimonious HEMC.[24]

The problem with this argument is that whatever ontological economy HEMC buys it does so at the cost of an increase in complexity elsewhere in the theory. We saw in Section 4 that an advocate of HEMC can employ explanations involving transcranial kinds, at the cost of introducing hybrid kinds into her theory. Psychological theories that use hybrid kinds involve neither wholly mental nor wholly non-mental predicates, they require a new kind of predicate that does not fit neatly on either side of the barrier: predicates that reference hybrid kinds. HEMC requires adding new predicates to the psychological theory and giving them a special treatment that differs from that of traditional psychologically-significant terms. This spares the ontology, but it complicates the theory. HEC avoids adding this wrinkle to the theory. Rather than add mental, non-mental, and hybrid predicates to the psychological theory, HEC allows one to take talk of psychologically-significant transcranial states (extended 'beliefs' and 'desires') at face value, assimilating them under the old categories of mental states. The choice is between adding new predicates to gain a sparser ontology, or beefing up one's ontology to include more mental 'stuff' while

24. Rupert (2004), p. 421: 'If the general [HEC] notion of access to information adds any explanatory power, it is too little to justify new ontological commitments.' Rupert (2009b): 'HEC explains the phenomena by positing the same number of elements, the internal architecture, the interactive process, etc., then lumps these parts together under the label "cognitive system". This addition is gratuitous.'

recycling the old predicates. The dilemma should be familiar to any metaphysician with naturalistic sympathies: ideological economy can be bought at the expense of ontological economy, and ontological inflation can be avoided at the expense of adding new primitive predicates. (Is it better to add new primitive modal operators to one's modal theory and dispense with possible worlds, or use possible worlds to buy a theory leaner in modal predicates?)[25]

The mere fact that there is a trade-off between ideology and ontology need not itself be a problem. The problem is that in the case of HEC and HEMC it is completely unclear how to manage the trade-off. Is it better to pare down ontology at the expense of mess in the psychological theory, or embrace an inflated ontology for a simpler treatment of psychologically-significant terms? The answer is unclear. Not only is it unclear, it is unclear how the practice of cognitive science would ever informatively settle this dispute. We appear, rather uncomfortably, not to be in a position to know how the explanatory merits of the two positions should be balanced. Consequently, there is no IBE from the parsimony of HEMC to the falsity of HEC. Parsimony has to be balanced against theoretical simplicity, and once one attempts to do this, Rupert's IBE against HEC stalls.

It is worth noting that in order for an IBE to fail, it is not necessary that there be *zero* difference in explanatory value between the two alternatives. All that is required is for there to be no clear winner. This appears to the case here: both HEC and HEMC can take on each other's explanatory work, but neither clearly, or *knowably*, trumps the other in terms of explanatory value. The explanatory race between them is too close to call, and it is hard to see how future evidence from cognitive science could settle it. A cognitive scientist could, perfectly rationally, adopt either framework for her day-to-day work. A persistent commitment to one framework over the other could be chalked up to individual prejudice, entrenchment of existing viewpoint, desire for different kinds of neatness, or an iconoclastic desire for revolutionary talk. None of these seem sufficient to warrant an inference to the truth or falsity of HEC. The conclusion is that IBE is the wrong tool to decide between HEC and HEMC. The explanatory practice of cognitive science does not have enough traction on the issue over which HEC and HEMC differ—the claim about mental extension—to mount an IBE.

Rupert alleges that HEC incurs an additional and different type of explanatory cost: HEC involves losing our grip of the human subject as a persisting and integrated system. Rupert claims that this means psychology misses out on perfectly good explanations of cognitive phenomena that are robustly exhibited across different environments: explanations in terms of the capacities of an invariant and integrated

25. For more on the interplay between ideology and ontology see Oliver (1996); Quine (1951).

human subject.[26] However, as we noted in Section 4, it is hard to see how this follows. HEC does not claim that we should abandon traditional psychology, or stop treating humans as persisting and integrated cognitive systems. It only claims that in addition to studying their internal mechanisms, we should also study larger mechanisms as units of cognitive activity in their own right. The mandate of HEC is to increase the explanatory scope of psychology, not to take old explanations away.[27]

6 Conclusion

The debate about the explanatory value of HEC to cognitive science is not about whether or not the mind extends. That issue is simply not sensitive to the explanatory practice of cognitive science. The debate *could* be about whether transcranial kinds should be allowed into cognitive science at all. Alternatively, it could be about whether the explanation of cognitive processes should be a purely internal matter (*à la* HINT). But on both scores, it seems that the externalist has already won. Transcranial kinds are already doing useful work in psychology as the studies of Gray and Ballard show. And psychology no longer assumes that cognition can wholly be explained in the internalist way envisaged by HINT. However, the externalist's prize claim—that the extra-neural material is *mental*—remains untouched. This is the claim that attracted the lion's share of the philosophical attention, and if it were dropped much of the heat would go out of the debate.

We can also conclude, *contra* Hurley, that appeal to the explanatory practice of cognitive science cannot settle the causal–constitutive dispute. Hurley proposed that the line between the non-mental causes of mental states and their mental constituents should be drawn based on the explanatory practice of cognitive science. We now have strong reasons to doubt that this strategy could succeed. Cognitive science is simply not sensitive to the difference between the two options. Psychological theories can, at negligible cost, be given either a causal gloss (via HEMC) or a constitutive gloss (via HEC). If, in a particular case, psychology chooses to favour one gloss, the other is still available via a trivial transform, and the reasons for a preference for one over the other appear to be more likely to be idiosyncratic and accidental rather than tied to tracking the truth. Explanatory practice cannot settle the issue between HEC and HEMC, *a forteriori* it cannot settle the causal–constitutive issue. Explanatory properties are toothless as a way of demarcating the causal–constitutive boundary.

26. Rupert (2004), pp. 425-428; Rupert (2008); Rupert (2009b).

27. See Clark (2007), pp. 169–171. Adams and Aizawa (2007); Aizawa (2007) develop a further IBE against HEC, based on Noë (2004), which I criticise in Sprevak (2009).

Does this mean that the choice between HEC and HEMC is radically underdetermined? Or worse, is the dispute between HEC and HEMC no more than a terminological question with no empirical bearing on the world? I do not think that the dispute is merely terminological. Interpreting the dispute between HEC and HEMC as merely a question of terminology fails to take substantial issues seriously. What is a mental process? Where are the cognitive agents? How far in space do they extend? The boundary between the mental and the non-mental is a natural joint, something that exists no matter how we choose to use our language. One cannot change facts about the distribution of mental states merely by legislating changes in our language. It is this ontological question that underlies the HEC/HEMC dispute. To say that the difference between a notebook being mental and non-mental is just a matter of the definition of the term 'mental' is to fail to hear this question seriously. If one were to insist that the mental/non-mental distinction, even in its ontological form, is nonetheless just a matter of definition, then one would have travelled far from the thought that mentality is a natural, objective, and real feature of the world.[28]

How then could we resolve the issue between HEC and HEMC? I believe that we need to look not just to which hypothesis is endorsed by scientific practice, but to wider considerations that inform what makes a state or a process mental or non-mental. In the case of cognitive processes, we already have theories of the mental/non-mental contrast in the form of various kinds of functionalism. These theories aim to specify what makes a state mental, at least for the non-qualitative, information-processing, aspects of our cognitive life. For example, a functionalist theory of belief may attempt specify what makes a causal state a belief. One might look to these theories of the mental/non-mental distinction to settle the HEC/HEMC dispute. In Sprevak (2009), I argue that rather than yielding a sensible answer, these theories tend to blow up when considering this particular problem. This does not show that we should give up and look elsewhere for answers, such as the explanatory practice of science. Rather, it shows that we need to rethink those theories carefully to avoid their failure in these cases. This process of remoulding our theories of mentality should be guided by more than just by the contingencies of the practice of cognitive science. There are wider concerns relevant to a theory of mentality that

28. A more reasonable reading of the terminology worry is to claim that our concepts of mentality, as they stand, are not adequate to the task of selecting between the two hypotheses. We need to revise our concepts to deal with the cases raised by cognitive extension, and perhaps the existing categories of mental and non-mental are too crude. I have considerable sympathy with this response, but it raises the problem of *how* we should reform our concepts of mentality. My point is that this cannot be just a matter of convention; one cannot simply legislate what counts as mental and what does not. Reform should be answerable to evidential concerns, and it is a puzzle where the evidence should come from if not from the practice of cognitive science. I argue that we should look wider for evidence, including our intuitions about merely possible cases.

have the potential to break a deadlock between HEC and HEMC. These include our judgements about merely possible cases of mentality and cognitive agents. The project of remoulding should be seen in the round, with empirical concerns informing in complex ways both intuitions about possible cases and the explanatory practice of cognitive science. The way forward is to acknowledge that whether something counts as mental depends on broader exigences than just the explanatory practice of cognitive science.

Clark says that 'the cure for cognitive hiccups [between HEC and HEMC] is to stop worrying and enjoy the ride'.[29] Clark claims that the interplay between HEC and HEMC has been, and is, a productive force in cognitive science 'apt to draw attention to certain features … while making it harder to spot others'.[30] However, this mistakes the genuine explanatory inferences driving the debate in cognitive science. The real rivals, as far as cognitive science is concerned, are not HEC and HEMC, or indeed any claims about the extent of the mental, but two different explanatory frameworks for understanding the mind: *internal self-sufficiency* (INT) and *external dependence* (EXT). It is the oscillation between these distinctively explanatory models that has been the force behind cognitive science:

INT Mental processes are largely self-sufficient, and can be studied largely in isolation from environmental props.

EXT Mental processes depend intimately on environmental resources, and should be studied within the context of those resources.

Some cognitive phenomena are best seen under INT, some are best seen under EXT. Understood right, the embedded turn in cognitive science has been a therapeutic series of hiccups between INT and EXT. However, in order to settle the question in which Clark is interested—the extent of mental states and processes—we have to look elsewhere for answers, and to our best theories of the mental/non-mental contrast.

Acknowledgements

This paper originated in a conversation with Peter Lipton after a lecture by Andy Clark in Cambridge. Andy gave a fascinating series of case studies in cognitive science that were intended to encourage one to believe in HEC. The inference appeared to be an IBE: HEC was presented as the best explanation of those case

29. Clark (2007), p. 192.
30. *ibid*.

studies. In the conversation, both Peter and I were drawn to the thought that this strategy would not work. Peter asked: 'What difference would it make if the environmental features are mental?' We struggled to find any. Sadly, our discussions were cut short by Peter's death. I hope that Peter would have been sympathetic with the general line of this paper, even if not with the details. It would have been much better for his input.

I was an undergraduate student of Peter, then a graduate student, and finally one of his colleagues at the Department of History and Philosophy of Science and King's College, Cambridge. Among Peter's many talents he possessed an uncanny ability, which could be given either a HEC or a HEMC parsing, to extend the mental abilities of those around him. Seminars, talks, supervisions, and conversations would be transformed by his presence, no matter what the topic. Often one did not know what one was thinking until it was refined, clarified, and infinitely improved by Peter ('Liptonised'). Working with Peter was both a friendly experience and engendered an almost magical sense of clarity. Like many others, I was looking forward to frequent future discussions and collaborations with Peter. Sadly, none of us will have this opportunity, and all our mental lives are considerably the worse for it.

I would like to thank Anjan Chakravartty, Andy Clark, Ken Aizawa, Zoe Drayson, Rob Rupert, and Mike Wheeler for helpful comments on an earlier draft of this paper. A version of this paper was presented at the ZiF conference in Bielefeld, Germany in November 2009.

References

Adams, F., and K. Aizawa. 2001. 'The bounds of cognition'. *Philosophical Psychology* 14:43–64.

———. 2007. *The Bounds of Cognition.* Oxford: Blackwell.

Aizawa, K. 2007. 'Understanding the embodiment of perception'. *The Journal of Philosophy* 106:5–25.

Ballard, D. H., M. M. Hayhoe and J. B. Pelz. 1995. 'Memory representations in natural tasks'. *Journal of Cognitive Neuroscience* 7:66–80.

Ballard, D. H., M. M. Hayhoe, P. Pook and R. Rao. 1997. 'Deictic codes for the embodiment of cognition'. *Behavioral and Brain Sciences* 20:723–767.

Burge, T. 1979. 'Individualism and the mental'. *Midwest Studies in Philosophy* 5:73–122.

———. 1986. 'Individualism and psychology'. *Philosophical Review* 95:3–45.

Chalmers, D. J. 2008. 'Foreword'. In *Supersizing the Mind,* 4–16. Oxford: Oxford University Press.

Chirimuuta, M., and I. Gold. 2009. 'The embedded neuron, the enactive field?' Chap. 9 in *The Oxford Handbook of Philosophy and Neuroscience,* edited by J. Bickle. Oxford: Oxford University Press.

Clark, A. 2005. 'Intrinsic content, active memory and the extended mind'. *Analysis* 65:1–11.

———. 2007. 'Curing cognitive hiccups: A defense of the extended mind'. *The Journal of Philosophy* 106:163–192.

———. 2008. *Supersizing the Mind.* Oxford: Oxford University Press.

———. 2009. 'Spreading the joy? Why the machinery of consciousness is (probably) still in the head'. *Mind* 118:963–993.

———. 2010. '*Memento*'s Revenge: The extended mind, extended'. In *The Extended Mind,* edited by R. Menary, 43–66. Cambridge, MA: MIT Press.

Clark, A., and D. J. Chalmers. 1998. 'The extended mind'. *Analysis* 58:7–19.

Darwin, C. 1962. *The Origin of Species.* 6th ed. New York, NY: Collier.

Fodor, J. A. 1980. 'Methodological solipsism considered as a research strategy in cognitive psychology'. (Reprinted in D. M. Rosenthal, editor, *The Nature of Mind*), *Behavioral and Brain Sciences* 3:63–109.

Fresnel, A. 1866. *Oeuvres Complètes.* Paris: Imprimerie Impériale.

Gray, W. D., and W.t.- Fu. 2004. 'Soft constraints in interactive behavior'. *Cognitive Science* 28:359–382.

Gray, W. D., C. R. Sims, W.t.- Fu and M. J. Schoelles. 2006. 'The soft constraint hypothesis: A rational analysis approach to research allocation for interactive behavior'. *Psychological Review* 113:461–482.

Hurley, S. 2010. 'Varieties of externalism'. In *The Extended Mind,* edited by R. Menary, 101–153. Cambridge, MA: MIT Press.

Lavoiser, A. 1862. *Oeuvres.* Paris: Imprimerie Impériale.

Lipton, P. 2004. *Inference to the Best Explanation.* 2nd ed. London: Routledge.

Marr, D. 1982. *Vision.* San Francisco, CA: W. H. Freeman.

McClamrock, R. 1995. *Existential Cognition.* Chicago, IL: Chicago University Press.

Noë, A. 2004. *Action in Perception.* Cambridge, MA: MIT Press.

Oliver, A. 1996. 'The Metaphysics of Properties'. *Mind* 105:1–80.

Putnam, H. 1975. 'The Meaning of "Meaning"'. In *Mind, Language and Reality, Philosophical Papers, vol. 2,* 215–271. Cambridge: Cambridge University Press.

Quine, W. V. O. 1951. 'Ontology and ideology'. *Philosophical Studies* 2:11–15.

Rupert, R. D. 2004. 'Challenges to the hypothesis of extended cognition'. *The Journal of Philosophy* 101:389–428.

———. 2008. 'Innateness and the situated mind'. In *The Cambridge Handbook of Situated Cognition,* edited by P. Robbins and M. Aydede, 96–116. Cambridge: Cambridge University Press.

———. 2009a. *Cognitive Systems and the Extended Mind.* Oxford: Oxford University Press.

———. 2009b. 'Keeping HEC in CHEC: On the priority of cognitive systems'. Manuscript.

Sprevak, M. 2009. 'Extended cognition and functionalism'. *The Journal of Philosophy* 106:503–527.

Thagard, P. R. 1978. 'The best explanation: Criteria for theory choice'. *The Journal of Philosophy* 75:76–92.

Van Fraassen, B. C. 1980. *The Scientific Image.* Oxford: Oxford University Press.

Williamson, T. 2000. *Knowledge and Its Limits.* Oxford University Press.