Published in *The Journal of Philosophy* (2009) 106: 503–527. mark.sprevak@ed.ac.uk

Extended cognition and functionalism

Mark Sprevak
University of Edinburgh

14 October 2009

Andy Clark and David Chalmers argue that the mind sometimes extends outside the body to encompass features of the environment (HEC). HEC has been criticised by Fred Adams, Kenneth Aizawa, and Robert Rupert. In this paper, I argue for two claims: (1) HEC is a harder target than those critics have supposed; HEC is entailed by functionalism, a commonly held view in philosophy of mind, and one to which those critics are already committed. (2) The version of HEC entailed by functionalism is more radical than the version that Clark and Chalmers suggest. I argue that this version of HEC is so radical as to form a counterexample to functionalism. The conclusion of the paper is against both HEC and functionalism.

Andy Clark and David Chalmers claim that cognitive processes can and do extend outside the head.[1] Call this the 'hypothesis of extended cognition' (HEC). HEC has been strongly criticised by Fred Adams, Ken Aizawa and Robert Rupert.[2] In this paper I argue for two claims. First, HEC is a harder target than Rupert, Adams and Aizawa have supposed. A widely-held view about the nature of the mind, functionalism—a view to which Rupert, Adams and Aizawa appear to subscribe—entails HEC. Either HEC is true, or functionalism is false. The relationship between functionalism and HEC goes beyond support for the relatively uncontroversial claim that it is logically or nomologically possible for cognition to extend (the 'can' part of HEC); functionalism entails that cognitive processes *do* extend in the actual world. Second, I argue that the version of HEC entailed by functionalism is more radical than the version that Clark and Chalmers suggest. I argue that it is so radical as to

1. Clark and Chalmers (1998).
2. Adams and Aizawa (2001, 2007); Rupert (2004).

form a counterexample to functionalism. If functionalism is modified to prevent these consequences, then HEC falls victim to Rupert, Adams and Aizawa's original criticism. An advocate of HEC has two choices: (1) accept functionalism and radical HEC; (2) give up HEC entirely. Clark and Chalmers' intermediate position of a modest form of HEC is unsustainable.

The argument of this paper, although initially appearing to support Clark and Chalmers, ultimately argues against their position. The price of HEC is rampant expansion of the mind into the world, and the implausibility of such expansion is indicative of deep-seated problems with functionalism. The argument of this paper consequently speaks to wider issues than just the status of HEC. The reasons for HEC's failure bring to light new troubles with functionalism as an account of cognitive systems.

In Sections 1–3, I give Clark and Chalmers' argument for HEC, Rupert, Adams and Aizawa's criticism, and my response. In Section 4, I argue that functionalism (of a minimal kind) entails HEC. In Section 5, I show that the modest version of HEC proposed by Clark and Chalmers is unsustainable. In Section 6, I analyse the features of functionalism responsible for generating radical HEC. In Section 7, I criticise the other main argument for HEC: that HEC should be accepted based on its explanatory value to cognitive science. I conclude that HEC, and the functionalism that supports it, should be rejected.

1 HEC

Clark and Chalmers introduce HEC with two thought experiments. The first thought experiment involves three ways of playing the computer game Tetris. In Tetris, the player rotates falling blocks to form complete horizontal rows which are then eliminated. Imagine:

T1 Sitting facing a computer screen and mentally rotating a block to judge whether it will fit the sockets below.

T2 Sitting facing a computer screen and physically rotating the image on screen by pressing a rotate button to judge whether the block will fit the sockets below.

T3 Choosing to perform the rotation using *either* old-fashioned mental rotation *or* a neural implant that quickly rotates one's mental image on demand.

First, Clark and Chalmers argue that the implant version of T3 is just as much a cognitive process as T1: there seems no reason why an implant cannot count as

cognitive merely because it is artificial, and one can imagine that the implant is as tightly integrated with the rest of the player's cognitive system as one likes. Second, they argue that T2 is just as much a cognitive process as T3. One can imagine that T2 and T3 have the same functional structure: the neural implant uses same algorithm for rotation as in T2, it is initiated in the same way (by motor cortex activity), and it produces output in a similar way (a retinal image). The difference is that in T2 the processing is spread between the agent and the computer, while in T3 all the processing takes place inside the agent. Since the question is whether cognitive processes *can* cross the skin/skull boundary, it would be question-begging to object that T2 is not cognitive only because it does cross that boundary. Clark and Chalmers claim that T2 and T3 are otherwise alike. Their conclusion is that T2 and T3 have an equal claim to be cognitive.

The second thought experiment involves dispositional belief. Inga hears of an exhibition at the Museum of Modern Art (MoMA). She thinks, recalls that MoMA is on 53rd St., and sets off. Otto suffers from a mild form of Alzheimer's and always writes down useful information in a notebook. He hears of the exhibition at MoMA, retrieves the address from his notebook, and sets off.

Clark and Chalmers claim that Otto's notebook plays a similar functional role to Inga's biological memory. The state of Otto's notebook interacts with Otto's desires and other beliefs in a similar way to the way in which Inga's biomemory interacts with her desires and other beliefs. Exposure to new information causes Otto to modify the state of his notebook. Exposure to new information causes Inga to modify her biomemory. The current state of Otto's notebook causes Otto to stop at 53rd St. The current state of Inga's biomemory causes Inga to stop at 53rd St. The functional role of the stored information—its 'functional poise'[3]—appears to be the same in both cases. Clark and Chalmers conclude that just as Inga has a belief that MoMA is on 53rd St., so Otto has a belief, with the same content, that extends partially into the environment.

Both cases rely on what Clark calls the 'parity' principle (which I will call the 'fair-treatment' principle). This principle guarantees equal treatment between internal and external cases. It states that if an extended process is relevantly similar to an internal cognitive processes (save for having external parts), then that extended process should have an equal claim to be cognitive. In short, one should not be prejudiced against extended processes. Extended processes should not have to meet a higher standard merely because they are extended.

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recog-

3. Clark (2010).

> nizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process.
>
> (Clark and Chalmers 1998, p. 8)[4]

The fair-treatment principle enables Clark and Chalmers to argue that if two processes are just like one another, save for one being internal and the other extended, then both have an equal right to be cognitive. The purpose of the Tetris and Otto/Inga cases is to show that, in the actual world, *there are* extended processes that are just like internal cognitive processes: Otto's notebook is functionally just like Inga's biomemory and T2 is functionally just like T3.

Rupert, Adams and Aizawa (RAA) accept the fair-treatment principle but reject Clark and Chalmers' treatment of the Tetris and Otto/Inga cases. RAA argue that once one considers the fine-grained functional structure of these cases, one can see that actual extended processes are *not* functionally like any internal cognitive process. The processes involved in T2 and Otto's notebook are so unlike any internal cognitive processes that they do not deserve to be called cognitive at all.

2 Criticism of HEC

Imagine memorizing a list of husbands and wives, A–B ('John–Mary', 'Peter–Jane', 'David–Sarah', etc.). Suppose that you are then told the couples divorce and remarry among themselves, and you attempt to memorize the new list of partners, A–C ('John–Sarah', 'Peter–Mary', 'David–Jane', etc.). Humans take significantly longer to learn the new A–C associations than the A–B associations, or a list of new associations. It appears that memory of the old A–B associations interferes with ability to acquire new A–C associations. This phenomenon is called 'negative transfer'.[5] Negative transfer is widely-exhibited in human memory (short-term memory, long-term memory, working memory, memory of names, stories, and numerical relations), but it is absent in the extended process described by Clark and Chalmers: it need not be any harder for Otto to write and recall A–C than it was for him to write and recall A–B.

If one were to adjust the Otto–notebook system so as to simulate negative transfer, there would be other features of human memory that the system would still lack: generation effects (better performance with self-generated mnemonics), satisfaction of power laws of remembering and forgetting, fit with human needs given the statistical properties of the environment, satisfaction of laws governing conditions of

4. See also Clark (2007).
5. Rupert (2004), pp. 413–415; Anderson (2000), pp. 239–243.

learning and extinction such as the Rescorla–Wagner law (Rupert 2004, pp. 416, 419). These features are characteristic of human memory, but not of writing information in a notebook. If one were to modify the Otto–notebook system to simulate *all* these features, then one would have moved so far from the original Otto–notebook scenario as to no longer have anything that corresponds to actual human tool-use. Therefore, Rupert concludes, extended memory processes do not occur in the actual world.

Adams and Aizawa argue similarly that the functional and causal structures of T1 and T2 are different. In T2, but not T1, there is muscle activity, and hence activation of motor processing systems. In T2, the agent must decide between two methods, which means that she must use additional cognitive systems: attentional mechanisms and memory to store the information that both methods are available. Finally, the causal structure of processes outside the agent in T2 (button pressing, electrons fired towards a phosphorescent screen) seem unlike the causal structure of processes that take place inside the head in T1.

In a similar vein, Adams and Aizawa argue that picking up the notebook and turning to an appropriate page requires use of Otto's motor systems; turning to an appropriate page and reading requires use of his visual systems; interacting with the notebook involves acquiring beliefs about the formal and physical nature of the notebook, e.g. that it is open at a particular page. These features are not reproduced in Inga's case. Therefore, the causal role of the notebook in Otto's cognitive life is not the same as that of belief in Inga's. The typical causes and effects of the notebook are so different from those of Inga's biomemory that there is no reason why if the latter is mental, we should think that the former is too.

3 Reply to RAA

Clark (2010) indicates a powerful response to RAA that I wish to elaborate. RAA argue that, on a fine-grained level, extended processes are functionally unlike internal cognitive processes and so do not deserve to be called cognitive. The response is that if one draws the boundary between the cognitive and the non-cognitive *this finely*, then one is committed to the claim that Martians cannot have cognitive processes.

The Martian intuition claims that it is possible for creatures with mental states to exist even if such creatures have a different physical and biological makeup to ourselves. An intelligent organism might contain green slime instead of neurons, it might be made out of silicon rather than carbon, it might have different kinds of connections in its 'nervous' system. There seems no reason why mentality has to involve blood, neural tissue, or DNA.

The Martian intuition applies to fine-grained psychology as well as physiology. There is no reason why an intelligent Martian should have exactly the same fine-grained psychology as ours. A Martian's pain response may not decay in exactly the same way as ours, its learning profiles and reaction times may not exactly match ours, the typical causes and effects of its mental states may not be exactly the same as ours, even the large-scale functional relationships between the Martian's cognitive systems (e.g. between its memory and perception) may not match ours.[6,7]

RAA focus on fine-grained features of cognition, such as negative transfer. But an intelligent Martian need not exhibit negative transfer. RAA focus on reaction-time patterns and learning curves. But a Martian need not exhibit the same reaction times or learning curves. RAA focus on characteristic errors. But a Martian need not make the same characteristic errors. RAA focus on Otto's use of his visual and motor systems to access his notebook. But a Martian may access its memory by using its visual and motor systems (it might store a memory by activating a certain pattern of motor activity, and it might retrieve a memory by seeing an image). RAA focus on additional attentional mechanisms used in T2 to perform rotation. But a Martian might use attentional mechanisms to perform mental rotation (it might deliberately decide between two different internal methods of mental rotation).[8]

One could imagine a Martian whose memory, instead of being stored in patterns of neural activity, was stored internally as a series of ink-marks. If the Martian wished to store new information, it would activate a process that would create new ink-marks in its storage system. If it wished to retrieve information, it would activate a process that would make a mental image of ink-marks appear in its visual system. It seems wrong to say that simply because a Martian stored its memories this way, we should deduce that it had no mental life, or lacked genuinely mental memory states or processes. In principle, there seems no bar to such a Martian having beliefs and mental states (provided, for example, it exhibited the relevant coarse-grained features of memory, and caused occurrent beliefs and desires in an appropriate way). Such a creature would have internal states with the causes and effects typical

6. The Martian intuition is typically formulated in the case of qualitative mental state (like pain). However, there is no reason why it should not apply to cognitive states (like belief), and cognitive processes (like inference).

7. Cf. Shoemaker (1984), p. 281: 'But what reason is there for thinking that these underlying processes and mechanisms [involved in perception, memory, information processing] must be the same in all creatures having mental states? In other words, what reason is there for thinking that all creatures having mental states must have the same "depth psychology"? As far as I can see, there is no reason for thinking that this is so, and there are good reasons for thinking that it is not.'

8. Note that the Martian intuition is not that a Martian could have a mental state that is in every aspect identical to a human mental state. The claim is that for a given *type* of human mental state (e.g. belief that *X*), it does not seem necessary to have human physical and fine-grained psychological makeup in order to have that state.

of the notebook in the Otto–notebook system. But just because a creature used ink-marks rather than neurons to store information, we would not conclude that it must thereby lack mental states.

RAA's objection to HEC is that fine-grained features of human cognition are necessary for mentality. But this seems wrong. Martians could differ from us in all kinds of fine-grained psychological ways and still have mental states. Therefore, such features are not necessary for mentality. This addresses RAA's worry, but it does not provide a positive argument for HEC. I wish now to argue that a number of varieties of functionalism entail HEC.

4 Functionalism entails HEC

Functionalism was in part designed to avoid necessarily withholding mentality from creatures with a different fine-grained makeup. Most versions of functionalism aim to save the Martian intuition in some form or other.[9] Functionalism preserves the Martian intuition by claiming that what makes an organism have a mental state is the organism's *functional organisation*. This is typically understood in terms of the notion of a *causal role*, which in turn is understood as a *pattern of typical causes and effects*. To a first approximation, one could describe the causal role of pain as follows:

> *Pain* is the state that tends to be *caused by* bodily injury, and *causes* the belief that something is wrong with the body and the desire to be out of that state; it also tends to *cause* anxiety, and, in the absence of any stronger, conflicting desires, wincing and moaning.
>
> (Levin, Fall 2004)

According to functionalism, any state that has this pattern of typical causes and effects is a pain state.[10] There is no reason why this state cannot be realized in a Martian with a silicon-based physiology and Martian fine-grained psychology, or in a human with a carbon-based physiology and human fine-grained psychology.

9. For versions that do not, see point 6 below.

10. Causal roles are usually described in terms of Ramsey sentences. This allows them to be specified without the use of mental state terms. The following Ramsey sentence roughly describes the causal role of pain: $\exists x \exists y \exists z$ (x tends to be caused by bodily injury & x tends to cause states y, z, a & x tends to cause wincing and moaning). No mental state terms appear in this sentence. If the sentence is expanded to include a theory of the causal relations between all our mental states, then it will (hopefully) specify all the appropriate causal roles informatively in a non-circular way. Although crucial to developing functionalism, the details of this method of specification make no difference to the current argument. See Lewis (1972) for the method.

Similarly, a Martian could have a belief state by having a state with the causes and effects typical of belief, or a cognitive process by having a causal process that relates its mental states in a way typical of that cognitive process.

Different brands of functionalism differ in how the causal roles should be specified. There are a number of dimensions of variation. The one to which I wish to draw attention is that the causal roles can be specified in *finer* or *coarser* grained detail. Adopting an appropriate level of detail is crucial to preserving the Martian intuition.

Human mental states have many typical causes and effects. As well as being impractical, it would be wrong for a functionalist to specify all the typical causes and effects. Some causes and effects are *ignored* in the functional specification. For example, a typical cause of pain in humans is apprehending a hurtful remark—ignored in the specification above. A typical effect of pain is decreased sensitivity to more minor injuries—also ignored. The reason some causes and effects should be ignored is that some causes and effects of human pain are not be essential to having pain. It is conceivable that a hurtful remark typically causes anger rather than pain in a Martian; and that pain causes an increase, or no change, in a Martian's level of sensitivity to more minor injuries.

While some causes and effects should be ignored, others should be *abstracted* to form a more general kind. All the following typically cause pain in humans: a blow to the head, a cut, a burn, and a gastric upset. Rather than enumerate each pain-causing event, a functionalist would do better to form the general kind *bodily injury*. This is not only more concise, it is also essential to preserving the Martian intuition. A Martian may not be able to suffer from gastric upsets, and it may be unaffected by, or feel pleasure from, a blow to the head.

All varieties of functionalism contain a parameter that controls how finely or coarsely functional roles should be specified (how much should be abstracted and ignored). If this parameter is set too fine, then one is committed to Martians who differ from us in minor ways not having mental states. If the parameter is set too coarse, then functional role specifications are too easy to satisfy, and systems that are intuitively non-mental wrongly count as mental. My claim is that if the grain parameter is set *at least* coarse enough to allow for intelligent Martians, then it also allows in many cases of extended cognition.[11]

11. The grain parameter is multi-dimensional since a functionalist theory needs to decide *which* features to abstract and ignore, not just a single magnitude *how much*. This does not affect the argument. Assume that grain parameter can vary along dimensions $\alpha, \beta, \gamma, \ldots$ This means that the grain parameter's components along $\alpha, \beta, \gamma, \ldots$ $(g_\alpha, g_\beta, g_\gamma, \ldots)$ need to be set coarse enough to allow Martians who differ along those dimensions to have mental states. But as described above, there are extended processes that depart from internal human cognition less jointly along each of those dimensions than possible Martians. Therefore, those extended systems also satisfy the functional roles.

The justification for the claim is that cases of extended cognition are *at least* as similar to cases of internal human cognition as possible Martian thought processes. Pick a putative case of extended cognition, e.g. the Otto–notebook system. One can imagine, as we did before, a Martian whose memory operates in the same way as Otto's notebook. Such a Martian's thought processes would be at least as different from internal human cognition as the Otto–notebook system—the Otto–notebook system at least contains Otto's internal human cognitive processes. However, we saw that just because such a Martian had a storage and recall system that operated in a different way from humans, that was no reason to conclude that it lacked genuine beliefs. If this is the case, then the functional roles associated with belief have to be set *coarse enough* to allow such Martians to have belief: creatures who might not exhibit negative transfer, and have different learning curves and reaction times. But if the functional roles are set *this* coarse, then they are also satisfied by the Otto–notebook system. Therefore, Otto's notebook counts as an extended belief.

Consider another example of extended cognition: counting on one's fingers.[12] One could imagine a Martian that uses inflation of fleshy tubes inside its head when counting. When the Martian wishes to add two numbers, it fills tubes in sequence (e.g. 2 + 3); another internal process detects how many tubes have been inflated. The mechanism could be initiated by activity in the Martian's motor system and yield output to its visual system. We would wish to say that such a Martian has a different mechanism for counting from us, but we would not wish to conclude that it must be non-cognitive as a result. Therefore, the functional roles characterising counting should be set coarse enough to allow such a Martian to satisfy them. But if they are set *this* coarse, then they are also satisfied by human–fingers systems.

The argument can be made in a stronger way. It is not hard to imagine intelligent Martians whose memory is *more* different from our own than Otto's notebook. A Martian might have a very bizarre way of storing and recalling memories: using ectoplasm or encoding memories using sub-atomic particles. Yet it seems possible that such creatures could nevertheless have beliefs. Therefore, a functionalist theory should set its grain parameter coarse enough not to rule out such creatures from having beliefs. But if it sets the grain parameter *this* coarse, then it will almost certainly be satisfied by the relatively modest departures from internal human cognition involved in the Otto–notebook system.

In short, if a functionalist theory sets the parameter that controls the level of grain at which two processes are functionally identical too fine, then intelligent Martians are not allowed, and HEC is false for the reasons that RAA suggest. If it sets this parameter too coarse, then intelligent Martians are allowed, but HEC comes out true. The problem that RAA face is that there is no intermediate setting of the

12. Clark and Chalmers (1998), p. 17.

parameter that: (i) allows preservation of the Martian intuition and (ii) makes HEC come out false. From a functionalist point of view, the mereological sum of us and our artefacts are actual Martians.

There are a few points to note.

First, I do not claim that there is a unique, or indeed any, correct setting of the grain parameter for functionalism. Different settings are appropriate for different kinds of functionalism, and someone hostile to functionalism might question if there is any reasonable setting at all. My claim is that no matter what setting is chosen, if it is sufficient to save the Martian intuition, then HEC comes out true.

Second, it is worth emphasising that the argument concerns the *actual* existence of extended cognitive processes, not their mere possibility. If functionalism admits possible intelligent Martians, then extended systems in the actual world also qualify as mental. One's attitude to non-actual Martians commits one to the truth of HEC in the actual world. A functionalist would be behaving like a NIMBY (not in my back yard) if she were to allow possible intelligent Martians, but not actual HEC.

Third, the argument is not specific to any particular brand of functionalism. It applies to any version of functionalism that saves the Martian intuition. If functionalism is understood, not in terms of typical causes and effects, but in terms of a rough-and-ready notion of functional organisation, a question of grain still arises: at what level of abstraction should one specify the functional organisation necessary for mentality? At too fine-grained a level, intelligent Martians are excluded. At too coarse-grained a level, intelligent Martians are allowed but so too is extended cognition. The same applies to Turing-machine functionalism, and functionalism based on explanatory, rather than causal, roles.

Fourth, the fine/coarse-grained distinction cross-cuts the scientific/folk-psychological distinction between functionalisms. Scientific functionalism (psychofunctionalism) looks to empirical science to specify the causal roles; folk-psychological functionalism looks to folk psychology. In both cases, a question of grain arises: how fine-grained should one specify the functional roles? Which (scientific/folk-psychological) causes and effects should one abstract or ignore? Both scientific functionalism and folk-psychological functionalism can save, or fail to save, the Martian intuition by specifying the causal roles in coarse or fine enough detail. If the causal roles are specified coarsely enough to allow intelligent Martians, then HEC comes out true.

Fifth, the argument is not specific to functional state identity theories or functional specification theories. David Lewis and David Armstrong's versions of functionalism allow mind–brain identity claims to be asserted.[13] On their view, the concept of pain

13. Armstrong (1968); Lewis (1983).

is a functional concept: anything that satisfies a certain functional role qualifies, by definition, as a pain state. Two physically type-distinct states 'pain'-in-Martians and pain-in-humans can qualify as genuine mental states, and as pain states, because both fall under the same mental concept, *pain*. It does not matter here whether Lewis and Armstrong are correct in this, note: (i) they too face a question about grain (which causes and effects are essential in the specification associated with mental state concepts?); and (ii) they aim to save the Martian intuition. Given (i) and (ii), once our mental concepts are specified coarsely enough to preserve the Martian intuition, they also admit HEC.

Sixth, not all versions of functionalism aim to save the Martian intuition. A psycho-functionalist might argue that the job of functionalism is only to capture *generalisations* concerning actual organisms, and functional roles need only be broadened enough to admit those creatures. This kind of functionalism eschews the Martian intuition, and so escapes the argument above. However, note: (I) Such a version of functionalism appeals only if one restricts attention capturing generalisations relevant to actual-world psychology. This is one job that a functionalist theory can perform, but it is also employed for a more metaphysical task: to give a solution to the mind–body problem. If one accepts this second application of functionalism then it is hard see why attention should be restricted only to actual organisms. (II) There may be enough variation between actual organisms that have mental states to broaden the functional roles sufficiently to make HEC true, at least for some claimed instances of extended cognition. Humans, chimpanzees, whales, dolphins, and octopuses have memory with different kinds of fine-grained characteristics. The differences in their functional architecture may push the grain parameter coarse enough to admit at least some cases of HEC. (III) There seems nothing to stop Martians from coming into existence, either naturally, or by deliberate construction on our part. What should we say on encountering such a creature? Would we say that it did not have mental states? Or would we broaden our characterisation of the functional roles to include it, and at the same time, allow HEC? If the latter, then why not admit *now* that such organisms, and hence actual extended systems, have mental states?

5 Radical HEC

Functionalism, if it saves the Martian intuition, entails HEC. However, it is unclear whether functionalism entails the version of HEC that Clark and Chalmers put forward. Clark and Chalmers add extra conditions to the functionalist credo:

H1 The [external] resource be reliably available and typically invoked.

H2 Any information thus retrieved be more-or-less automatically endorsed. It should not usually be subject to critical scrutiny (unlike the opinions of other people, for example). It should be deemed about as trustworthy as something retrieved clearly from biological memory.

H3 Information contained in the resource should be easily accessible as and when required. (Clark 2010, pp. 6–7)[14]

What justifies these extra conditions? Clark and Chalmers say nothing except that H1–H3 make HEC more modest and more plausible. The problem is that this modest form of HEC is incompatible with a functionalist defence of HEC. Consider the conditions one at a time.

H1: For a resource to be cognitive, it does not seem necessary that it be reliably available or typically invoked. One could imagine a Martian with internal cognitive resources that are neither reliably available nor typically invoked. The Martian might have cognitive resources that are only available after it gets a good night's sleep, and it does not reliably or often get a good night's sleep. However, that does not stop, on those occasions when the Martian does get a good night's sleep, from those resources counting as genuinely cognitive. Another example is that the cognitive resources involved in acts of outstanding human creativity are not reliably available or typically invoked, but if someone does employ them, then that activity counts as part of their cognitive process. Just because a cognitive process is not reliably available or typically invoked, that does not make it non-cognitive or non-mental.

The same argument as Section 4 can be run. If the functional roles of cognitive states and processes are specified broadly enough to allow for *internal* resources not to be reliably available or typically invoked, then they should allow *external* resources not to be reliably available or typically invoked either. Special pleading for constraints on the external but not the internal conflicts with the assumption—that conditions that favour the internal over the external should be argued for, not stipulated *ad hoc*—on which the argument for HEC was based.[15]

H2 requires that information retrieved from an external cognitive resource be: (i) more-or-less automatically endorsed; (ii) not subject to critical scrutiny; (iii) deemed as trustworthy as that from biological memory. All conditions are violated by actual and possible cases of internal cognition.

14. Clark and Chalmers (1998), p. 17.

15. If H1 is modified to the condition that under *ideal circumstances* the resource should be reliably available and typically invoked, then a related problem arises: justify why ideal circumstances should not include the presence of those external objects that are rarely or unreliably available, and do so in a way that does not beg the question against HEC, i.e. that does not exclude appeal to them simply because they are external.

One could imagine a creature whose internal resources violate H2/i. A Martian might (redundantly) run several cognitive processes on a problem, compare the results, and only endorse a result if all (or most) agree. Just because the Martian adopts a cautious attitude towards its cognitive processes, that does not make those processes non-cognitive or non-mental. Some human internal cognitive processes do not have their output automatically endorsed either, e.g. imagining, supposing, desiring. If internal cognitive process do not satisfy H2/i, then why should extended processes? Condition H2/ii fails for a similar reason: a cautious Martian does routinely subject the output of its cognitive resources to critical scrutiny, but its cognitive resources are not made non-cognitive or non-mental as a result. Condition H2/iii also fails. One could imagine a Martian with internal memory less trustworthy than human internal memory, but still trustworthy enough to count as memory. One could imagine a series of creatures whose internal storage and recall mechanisms are, in different ways, less trustworthy than internal human memory but still trustworthy enough to count as memory. If internal resources can be less trustworthy than biological memory, then why not external resources? Moreover, as was indicated for H2/i, there are actual human internal cognitive resources that contain information that is not deemed trustworthy at all.

H3 requires that information present in the external resource be easily accessible as and when required. This condition is also violated. A Martian might have information in its internal resources that it finds difficult to access, e.g. it might have beliefs that it finds difficult to access. The Martian might need help, such as talking to a psychotherapist, to access some of its buried beliefs. But just because some of its beliefs are difficult to access, that does not make them less mental or cognitive. Humans also have mental information stored in internal cognitive resources that cannot be easily accessed. The visual system contains information about current eye position that cannot be easily accessed. Conscious beliefs can also be difficult to access. A nervous student might cram information into her cognitive resources before an exam that she is too nervous to recall during the exam, and subsequently forgets, but although the information was never *easily* accessible, that did not stop it from counting as genuinely cognitive while she had it.[16,17]

16. Again, appeal to normal or ideal conditions does not help. It raises the problem of justifying why normal or ideal conditions should not include the easy accessibility of information in external artefacts that are typically inaccessible.

17. Clark and Chalmers (1998) propose a fourth condition:

H4 The [external] resource be reliably available and typically invoked.

As Clark and Chalmers admit, H4 is false of internal human cognition: one can acquire beliefs through subliminal perception or memory tampering. It is also possible that a Martian could have innate beliefs that it did not previously consciously endorse. Another problem is that H4 only applies to memory and not to other cognitive processes.

Let us rehearse why H1–H3 should be rejected. H1–H3 are violated by actual and possible cases of internal cognitive resources. This creates a problem for HEC: why, if internal resources are allowed to violate H1–H3, should external resources not be? What justifies the differential treatment? Why require that extended processes meet a higher bar? If differential treatment is acceptable here, then why not at the beginning of the argument for HEC when it was claimed to be unacceptable to treat a resource differently simply because it was external? The problem with H1–H3 is that they violate the fair-treatment principle. The fair-treatment principle requires that if an external case is judged non-cognitive, then it should not be simply because it is external (equivalently, different standards should not apply simply because it is external). External and internal cases should be treated in an even-handed manner: if an extended process is relevantly similar to an internal process, save for having external parts, then it has an equal claim to be cognitive. The fair-treatment principle is required in order to make the functionalist argument for HEC work. If an advocate of HEC violates this principle, then she blocks her own argument for HEC.

Consequently, functionalism entails HEC, but not the modest version of HEC that Clark and Chalmers put forward. However, the functionalist argument for HEC does entail a radical form of HEC: HEC unqualified by extra conditions. The problem is that this radical form of HEC is almost certainly false. It is wildly over-permissive in attributing mental states. Here are some examples.

According to radical HEC, simply by picking up a book, I come to believe everything contained in that book. The justification is as follows. Consider a Martian like the one discussed in Section 4 who encodes memories using ink-marks. As well as acquiring beliefs via its senses, it seems possible for such a Martian to be born with innate beliefs. Furthermore, it seems possible for an organism to have innate beliefs that it has not examined yet—a library of data that is hard-wired into the organism by developmental processes, which the organism has not yet had cause to employ. Imagine that an ink-mark-based Martian is born with a stock of innate beliefs, most of which it has not chosen, or had cause, to examine yet, but it could if it wanted to. It seems conceivable that such Martian could exist. The Martian has ink-marks inside its head that, if it were sufficiently diligent, would guide its action in appropriate ways; I have ink-marks just outside my head that, if I am sufficiently diligent, would guide my actions in appropriate ways. The difference between the Martian and me is that it has the ink-marks inside its head, while I have the ink-marks outside. By the fair-treatment principle, if the Martian has the beliefs, then so do I.

The same argument applies to cognitive processes. Imagine that my desktop computer contains a program that calculates the dates of the Mayan calendar 5,000 years into the future. As a matter of fact, I never run this program, entertain the

question of what the Mayan calendar is for any year, or even know that my computer contains such a program. However, if I wanted to know the Mayan calendar and explored the resources of my computer, the program would allow me to find the answer quickly. According to the functionalist argument above, I possess a mental process that calculates the dates of the Mayan calendar. The justification: one could imagine a Martian with an internal cognitive process that calculates the dates of the Mayan calendar using the same algorithm. The Martian's ability could be innately present as an unintended by-product of the unfolding of its genetic program. The Martian may never happen to use this cognitive process; it may even be unaware that it has this cognitive process. However, like the card-counting of Raymond Babbitt in *Rain Man*, the Martian may find such a cognitive capacity awakened under the right circumstances, and that it can easily answer relevant questions. The Martian would be deploying an underused and hitherto dormant cognitive process. By the fair-treatment principle, I also have that cognitive process (and similarly for other programs of which I am unaware on my computer).[18]

Another example: Abel is a calculating prodigy who can perform feats of mental arithmetic. Baker is a normal human subject equipped with a supercomputer. The functionalist argument wrongly entails that Baker's mental arithmetic powers outstrip Abel's. The justification: it is possible for a Martian to exist with the same internal functional organisation as the joint Baker–supercomputer system. Therefore, by the fair-treatment principle, Baker's interactions with the supercomputer count as part of his cognitive process. Furthermore, one cannot say that what Abel does is 'more mental' than Baker. Both have an equal claim to mentality. The only difference is that their mental processes have a different fine-grained structure. However, it seems plain wrong to say what Baker does is just as mental as Abel.

In order to rule out these cases it does no good to say that the relevant extended systems only exist intermittently or for short periods of time. It is possible to imagine the relevant populations of Martians popping in and out of existence intermittently and for short periods of time without their internal processes being made non-mental as a result. Similarly, as Clark and Chalmers argue, even if a neural implant is only plugged in occasionally that need not stop it from counting as part of one's mental activity when present.[19]

18. It is no objection that my activation of the computer program may require intentional action on my part. One could imagine that the Martian requires similar intentional action, e.g. conscious searching through its internal cognitive resources, in order to waken its dormant cognitive process. Moreover, it is not clear that intentional action on my part is even necessary to parallel the Martian case—one could imagine that the Mayan-calendar computer program just happens to be launched in a fortuitous set of circumstances; once it is active I can easily answer questions about the Mayan calendar.

19. Clark and Chalmers (1998), p. 11.

The examples can be elaborated. By considering appropriate Martian scenarios, one can argue that if I step into a library, I instantaneously acquire millions of beliefs. By browsing the internet, I instantaneously acquire billions of beliefs. If we swap our address book, we instantaneously swap our beliefs. Although human memory is not like a library, the internet, or an address book—these have been abandoned as psychological models—it is *conceivable* for an intelligent being to have a memory resource that does operate in that way. Those psychological theories may be false of humans, but they are not (or at least, not obviously) incoherent. This mere possibility is enough for the argument in Section 4 to work.

These consequences of radical HEC seem false. Radical HEC should therefore be rejected. Is there a more modest form of HEC that is acceptable? We saw that adding H1–H3 to tame HEC did not work. Are there other conditions that can be added without disrupting the functionalist argument?

There are two reasons why such conditions are unlikely to be found.

First, any such conditions would have to satisfy the fair-treatment principle: they should be satisfied not just by actual extended cognitive systems, but also by all actual and possible internal cognitive processes. But given the *vast* variety of possible internal cognitive processes, such a condition would hardly be any constraint at all. The argument can be phrased as follows. For any instance of actual human tool-use p for manipulating representations, one can imagine a Martian who is otherwise like us, but with p as one of its internal processes. It seems perfectly coherent for p to count as one of the Martian's cognitive processes—just imagine an organism identical to us, but with some extra cognitive abilities or quirks involved in having internalised p. If p qualifies as a possible internal cognitive process, then it cannot be ruled as non-cognitive by any extra conditions: that would violate the fair-treatment principle. It would be to say that if p were to occur internally it would be cognitive, but when p occurs externally in an otherwise functionally identical system it is non-cognitive. If p cannot be excluded by the extra conditions, then it can only be excluded on functionalist grounds, and we have already shown that those are too weak. Therefore, if an extra constraint has to satisfy the fair-treatment principle, it is hard to see how it can be any substantial constraint at all.

Second, it is not clear how adding an extra constraint would help to avoid radical HEC anyway. Adding an extra constraint does not, by itself, disrupt the plausibility of the Martian scenarios that generated radical HEC. If one admits that the Martians described above have beliefs and cognitive processes, then by the functionalist argument and fair-treatment principle, so do the corresponding extended systems. Adding an extra condition does not block this inference. At most, it excludes further internal–external functionally equivalent pairs from counting as cognitive. The only way to avoid radical HEC is either: (i) drop the fair-treatment principle, or (ii) drop

the claim that the Martians in those cases have mental states. The first option is unacceptable as a way to defend modest HEC. The second option is unpromising too. If one were to give up the Martian intuition entirely, then RAA's criticism returns. If one wishes to save *just* those Martians that yield a modest form of HEC, then the question arises of what makes this more than an *ad hoc* manoeuvre to make modest HEC true. Why should mentality be granted to exactly those Martians but not others? Why save an ink-mark-using Martian without innate beliefs, but not an ink-mark-using Martian with innate beliefs? What justification, other than the truth of modest HEC, is there to restrict mentality to just those Martians?

Another option for defending modest HEC is to say that H1–H3 should be kept, not as individually necessary conditions for cognition, but along with the familiar functionalist condition, as jointly sufficient. Call such a theory JS-HEC. The Otto–notebook system satisfies JS-HEC and so counts as mental. Paradigm internal instances of cognition satisfy JS-HEC and so count as mental. However, radical instances of extended cognition do not satisfy JS-HEC. JS-HEC is silent about these cases: it is an incomplete theory of mentality, but one that supports modest HEC.

However, such a manoeuvre only postpones problems. The fair-treatment principle pushes JS-HEC beyond silence about functionally equivalent cases. The fair-treatment principle requires that if the only significant difference between two processes is that one is extended and the other internal, then both should have an equal claim to mentality. If two processes have an equal claim to mentality then it would be disingenuous, and at worst misleading, to claim that one is mental while remaining silent about the other. The examples of radical extended cognition described above do not satisfy JS-HEC, and hence JS-HEC does not judge them to be mental. However, they are functionally similar to internal cases that *do* satisfy JS-HEC. Therefore, by fair-treatment the external cases should count as mental too: if those processes were to take place inside the head, then we would call them cognitive. Therefore, JS-HEC plus fair-treatment inflates to radical HEC.

A variant of JS-HEC is to restrict functionalism's application to only those internal processes that do not have external counterparts that generate radical HEC. A non-functionalist theory would be given for other internal processes, or we could remain silent about them. Here, the fair-treatment principle would not get a grip, because the internal cases corresponding to radical HEC are not even considered. The problem is that this response, like that of selectively saving only those Martians that yield a modest form of HEC, appears to be an unacceptably *ad hoc* way to defend modest HEC. What justification is there for *this* particular division in the treatment of internal processes? What reason is there to restrict functionalism to only those internal processes that yield modest HEC and no more? There seems no reason other than a question-begging fondness to save modest HEC.

We have seen that the relationship between functionalism and HEC is an intimate one: functionalism that saves the Martian intuition entails HEC. This appeared to give us good reasons to think that HEC is true. However, functionalism only entails a radical form of HEC. This form of HEC violates so many pre-theoretical intuitions about mentality, that it is evidently false. The connection between functionalism and HEC now works the other way: if functionalism only entails radical HEC, and radical HEC is false, then functionalism is also false. Rather than HEC being a surprising true consequence of functionalism, it is a counterexample to that theory.

6 The problem with functionalism

Which aspects of functionalism generate this problem? The fault appears to lie in the joint acceptance of the following two intuitively plausible principles:

F1 If an organism counts as sufficiently like us on a coarse-grained global functional comparison, then it is a cognitive agent.

F2 If a cognitive agent contains a representation-manipulating process that is significant for guiding its action (in appropriate ways) when employed, then that process is one of its cognitive processes.[20]

F1 is plausible on straightforward functionalist grounds. F2 is plausible on the grounds that there could be *alien* cognitive processes: cognitive processes that are not similar, on any piece-wise comparison, to any actual internal human cognitive process. Martians might have different sensory modalities, and different ways of processing them. It would be chauvinist to exclude such processes from mentality because they do not have human equivalents. To a first approximation, it appears that our intuitions about alien cognitive processes are guided by something like F2: if a system already qualifies as a cognitive agent, then we are relatively generous in allowing its representation-manipulating processes to count as cognitive processes. The key judgement is the global judgement: decide if an organism is sufficiently like us to qualify as a cognitive agent. Once this judgement has been made, all kinds of internal processes can count as cognitive.

F1 and F2 generate a wide range of Martian scenarios. Once the fair-treatment principle is added, everything is in place for the inference from the possibility of

20. The 'appropriate way' clause is to handle Block (1978)'s elementary particle people, who have representations inside their heads but those representations do not appear to guide the overall agent's action in a way that qualifies as one of the agent's cognitive processes. This clause does not effect the argument below.

intelligent Martians to radical HEC. The fault with this inference seems to lie in either F1, F2, or fair-treatment. The difficulty is that it is not clear how any of these principles can be rejected.

An initially plausible target is F2. But an attack on F2 raises the question of how to distinguish alien processes that do deserve mentality from those that do not. F2 may get the details wrong, but something with a similarly liberal nature seems to be correct. Once we have judged that an agent is cognitive, we appear to be extremely permissive in allowing representational processes inside that agent to count as cognitive. So if F2 is rejected, something still liberal enough to support radical HEC is likely to take its place. If one rejects not just F2, but also those more general liberal intuitions, then a worry about chauvinism arises: Why cannot Martians have extra sensory modalities? Why cannot Martians have unique cognitive processes? Cannot one make sense of humans having extra, or different, cognitive processes?

F1 has already received a great deal of attention. It does not seem open to anything other than a wholehearted rejection of functionalism. As we saw in Section 4, F1 cannot be qualified by adjusting the grain parameter to allow for intelligent Martians but not extended cognition. While the agent has a tool to hand, the joint agent–tool system qualifies, in terms of a coarse-grained functional comparison, as a cognitive agent.

The fair-treatment principle is also hard to reject. One might try to weaken the principle to allow for the *possibility* of extended cognition, but not require that internal and external cases be treated *equally*. It is conceivable that internal processes could carry more weight in judgements of mentality than extended processes. This raises the question of how much extra weight internal cases should carry. If internal cases carry too much weight, then any argument for HEC disappears. If internal and external cases carry equal weight, then radical HEC results. What intermediate setting should be used? Again, we appear to have nothing to guide us in our decision other than brute, contested, intuitions about the truth and falsity of modest HEC. But it would be question begging to modify fair-treatment just enough to yield modest HEC, justified by the intuition that modest HEC is true, while going on to employ that principle as an argument for modest HEC. In any case, it is worth noting that the fair-treatment principle seems plausible as it stands.

Neither F1, F2, nor fair-treatment are obvious candidates for rejection. A more radical approach may be in order. The correct lesson may be that our intuitions about mental systems cannot be systematised without doing serious damage to our concept of mentality. Functionalism aims to provide an answer to what makes certain systems mental. Perhaps such an answer cannot be given. There may be no single specification, either physical or functional, that all and only mental states and processes of a given kind satisfy. The most we can say is that competent observers

agree that some systems and processes are mental and others are not, and in some cases no agreement can be reached. There is no underlying theory to be given of the mental/non-mental contrast. Mental systems do not form a natural kind.

This kind of quietism about mentality is no easy resting place for two reasons. First, the mental/non-mental contrast seems like a genuine joint, and the thought that some unified account of it can be given is still compelling. Second, on the quietist view we not only lose any argument for HEC, but also any argument against HEC. The hope was that one could appeal to general theories of mentality to decide whether extended cases were mental or not. If quietism is correct, then there is no way of resolving these cases: they are simply cases where competent observers differ.

7 Metaphysical vs. explanatory arguments for HEC

Clark and Chalmers have a second argument for HEC that they do not clearly distinguish from the metaphysical argument above. This argument is in terms of HEC's explanatory value. Their claim is that, not only are extended processes metaphysically like cognitive processes, but also that it *suits the explanatory aims of cognitive science* to treat extended processes as cognitive processes. The explanatory value of HEC to cognitive science is an argument for HEC's truth. In contrast, RAA argue that it suits the explanatory aims of cognitive science *not* to treat extended processes as cognitive processes, and therefore we have good reasons to reject HEC. Hence, there are good reasons to think that HEC is false.[21]

Both Clark and Chalmers and RAA think that the explanatory value of HEC to cognitive science is a guide to HEC's truth. Their disagreement concerns whether HEC's explanatory contribution is positive or negative. I wish to argue that both Clark and Chalmers and RAA are mistaken on this point. Any attempt to settle the status of HEC by appeal to explanatory value to cognitive science is misguided. There is no inference from the explanatory value of HEC to its truth/falsity, because a competing hypothesis exists with (almost) the same explanatory value but a

21. Rupert: 'HEC's plausibility depends on …[it providing] a coherent and fruitful framework within which to place all, or at least a healthy majority of, significant results in cognitive science …If the cases canvassed here are any indication, adopting HEC results in a significant loss of explanatory power or, 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, pp. 407, 390). Adams and Aizawa: 'In contrast to intracranial processes, transcranial processes are not likely to give rise to interesting scientific regularities …There just isn't going to be a science covering the motley collection of "memory" processes found in human tool use …Our view is that cognition$_{cs}$ [without HEC] will produce a natural science, where cognition$_{c}$ [with HEC] will not' (Adams and Aizawa 2001, pp. 61–62).

divergent truth value from HEC. The failure is typical in inference to the best explanation: the existence of a serious competing alternative.

The alternative hypothesis, the hypothesis of embedded cognition (HEMC), was introduced by Rupert, but he does not seem to acknowledge that it undermines his argument for the falsity of HEC as well as Clark and Chalmers' argument for its truth. HEMC, like HEC, claims that the study of cognition should involve an understanding of how an agent exploits its environment. HEMC acknowledges that cognitive processes depend in just the way that HEC suggests—intimately, and in hitherto unexpected ways—on the presence of external props and the structure of the environment. However, HEMC stops short of claiming that those external props are mental. According to HEMC, extra-cranial features play an essential role in cognition; according to HEC, those features play the same role and they are mental. In short, HEMC is HEC shorn of the claim that the extra-cranial features are mental.

The difference in explanatory value to cognitive science between HEMC and HEC is small. Both advocate the same kind of reform of cognitive science to include the study of mind–world relationships. Similar explanations are available in each case. It is hard to imagine any cognitive phenomenon that HEC, but not HEMC, can explain or vice versa.[22] A working cognitive scientist could switch between the two frameworks with little or no modification of her empirical work. The turn from individualism to embedded cognition is radical, but once that turn has been made, there is little to choose, in terms of explanatory value to cognitive science, between the two frameworks.

There is little to choose, but Rupert claims that HEMC is 'more conservative' and hence should be preferred.[23] However, this is unclear. HEMC does not claim that as much of the world is mental, but HEC is more conservative along a different dimension: it requires fewer steps in the explanation of action. On HEC, one can explain why Otto walked to 53rd Street simply by saying that Otto wanted to go to MoMA and believed that it was on 53rd Street:

> The alternative is to explain Otto's action in terms of his occurrent desire to go to the museum, his standing belief that the Museum is on the location written in the notebook, and the accessible fact that the notebook says the Museum is on 53rd Street; but this complicates the explanation unnecessarily … to explain things this way is to take *one step too many*. It is pointlessly complex, in the same way that it would be

22. Rupert provides a detailed description of how HEMC can replace HEC in cognitive science with little or no explanatory loss illustrated with a number of examples. I will not repeat these here. Rupert develops HEMC from McClamrock (1995).

23. Rupert (2004), pp. 395, 405, 421, 424.

> pointlessly complex to explain Inga's actions in terms of beliefs about her memory. The notebook is a constant for Otto, in the same way that memory is a constant for Inga; to point to it in every belief/desire explanation would be redundant. In an explanation, simplicity is power.
>
> (Clark and Chalmers 1998, p. 13)

Therefore, HEC is not straightforwardly explanatorily poorer than HEMC. It posits more mental activity, but it has the virtue of allowing cognitive science to give shorter explanations.

HEC and HEMC have slightly different explanatory shapes. The question is whether their different explanatory shapes yield a net explanatory value sufficiently, and knowably, different to warrant an inference to the best explanation. Are they different enough that we can say that one is clearly better than the other? *Contra* Clark, Chalmers, and RAA, the differences provide no argument for the truth or falsity of HEC. This is for two reasons.

First, although their explanatory properties are different, the net gains and losses in moving from one hypothesis to the other appear to be relatively minor and arguably negligible considering the wider explanatory aims of cognitive science. The explanatory gains and losses described above do not dramatically further the aims of cognitive science; at best, they amount to small tweaks around the edges. A cognitive scientist could, perfectly rationally, prefer one framework over the other. A persistent commitment to one framework 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.

Second, even if one does think that the gains and losses are explanatorily significant, it is far from obvious which hypothesis would win in a trade-off, or whether there would be a uniform winner in all cases of psychological explanation. We are not in a position to know how the explanatory costs of the respective positions should be balanced (positing more mental stuff vs. longer explanations). For an inference to the best explanation to fail, it is not necessary that there be zero difference in explanatory value between the competing alternatives. All that is required is that there be no clear winner. This seems to be the case, as the possibility of rationally preferring one explanation to the other appears to suggest. Notably, this is not a matter of lack of empirical knowledge. No matter how much empirical knowledge

we acquire, the explanatory winner would still not be clear.[24,25]

Therefore, although HEC and HEMC have different explanatory characters, to the best of our knowledge, their net explanatory worth is not significantly different. This invalidates both inferences to the best explanation. There is no inference from the explanatory value of HEC to its truth, because HEMC, to the best of our knowledge, has no less explanatory value than HEC and holds HEC is false. There is no inference from the lack of explanatory value of HEC to HEMC's truth and HEC's falsity, since HEC, to the best of our knowledge, is not significantly explanatorily worse off than HEMC. It is of course open to attack *both* HEC and HEMC, but this is not an option that Rupert or Adams and Aizawa wish to pursue. Neither wish to deny the externalism common to both hypotheses. They deny only the extra metaphysical claim made by HEC.

One might compare the failure of the explanatory argument for HEC with similar failures in other problematic cases for functionalism. Although it is controversial whether Ned Block's China-headed robot would or would not have mentality, the question cannot be settled by explanatory value to cognitive science. It would make little explanatory difference to cognitive science either way, even if such a robot were actual. Similarly, whether a tool is literally part of our mind or merely an essential non-mental prop makes little difference to cognitive science. The metaphysical difference between HEC and HEMC has vanishingly little traction on the day-to-day work of cognitive science. Awarding the encomium 'mental' may have rhetorical value in focusing attention on previously ignored environmental features. But appeal to rhetorical value is no argument for the truth of HEC.

24. Clark (2007) claims that empirical evidence favours HEC over HEMC. In reply to Rupert, Clark cites processes for which empirical psychologists find it fruitful to consider recruitment of extra-cranial resources (see Goldin-Meadow (2003); Gray et al. (2006); Paul (2006)). However, Clark does nothing to establish the crucial point that any clear explanatory benefit accrues to the claim that these extra-cranial resources are mental, as opposed to essential non-mental props in intimate law-like relations with the agent (pp. 171–174, 183–189).

25. Aizawa (2007) claims that scientific evidence favours HEMC over HEC in at least one case. The case is Noë (2004)'s claim that perceptual experience is constituted by sensorimotor skills (COH) as opposed to merely being causally related to sensorimotor skills (CAH). Aizawa claims that CAH should be preferred to COH because it better accounts for the fact that neuromuscular blockades paralyse sensorimotor skills but leave perceptual experience intact. Suppose that Aizawa is right—perception should be regarded as an essentially different department of the mind from sensorimotor skills. This settles an aspect of cognitive architecture, but does not settle the HEC/HEMC question. Are extra-cranial sensorimotor skills mental or are they merely non-mental adjuncts intimately related to the central cognitive resources of the agent? Both hypotheses are compatible with Aizawa's data about paralysis. The problem is that COH makes a claim about both cognitive architecture and HEC. Aizawa's evidence only tells against the former. Similarly, Noë's empirical evidence supposedly in favour of COH is compatible with HEC and HEMC (Aizawa 2007, pp. 11–20).

8 Conclusion

The most plausible justification of HEC is the functionalist argument given in Section 4. The Martian intuition is central to this argument. It is perhaps surprising that the existence of actual extended minds turns out to depend on our attitude towards possible Martians. The Martian intuition is two-edged however: it provides a defence against RAA and an argument for HEC but it commits one to a radical form of HEC. It is imaginable that someone might stubbornly assert the truth of radical HEC in the face of this argument. However, she would win few friends by doing so. Although not overtly contradictory, radical HEC violates so many pre-theoretical intuitions that it simply appears to get the facts about mentality wrong. If one is insensitive to this, then it is unclear how any evidence, short discovery of outright contradiction, could bear against radical HEC. Unless one wishes to dogmatically maintain HEC come what may, radical HEC, and the functionalism that supports it, seem false.

HEC is still compelling as a metaphor. It effectively spurs one's attention to mind–world relationships. This is perhaps where the real value of HEC lies: not as a claim that we have reason to believe is true, but as a claim that serves a rhetorical and heuristic purpose for cognitive science. It draws out attention to mind–world relationships, and it dramatises their importance. A second purpose for HEC is that it can serve as a constraint on theories of mentality. Abstracting away from the details of functionalism, it is hard to come up with any theory of mentality that allows for possible intelligent Martians but avoids false claims about radical extended cognition. Human–artefact interactions can be added to the familiar array of test cases for a theory of mentality. HEC, indicative of problems with functionalism, may be helpful in shaping a successor theory.

Acknowledgements

Thanks to Michela Massimi, Fred Adams, Arif Ahmed, Ken Aizawa, David Chalmers, Andy Clark, Richard Harper, Jane Heal, Peter Lipton, Matteo Mameli, Hugh Mellor, Matthew Nudds, Michael Potter, Nicholas Shea, Zoltán Szabó, and Mike Wheeler for comments and suggestions on earlier versions of this paper. Part of this paper was presented at the European Society for Philosophy and Psychology, Geneva, 2007. I would like to thank Microsoft Research for funding this work.

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.

Anderson, J. R. 2000. *Learning and Memory.* New York, NY: Wiley & Sons.

Armstrong, D. M. 1968. *A Materialist Theory of the Mind.* London: Routledge.

Block, N. 1978. 'Troubles with Functionalism'. In *Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science,* edited by C. W. Savage. Minneapolis: University of Minnesota Press.

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

———. 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.

Goldin-Meadow, S. 2003. *Hearing Gesture: How Our Hands Help Us Think.* Cambridge, MA: Harvard University Press.

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.

Levin, J. Fall 2004. 'Functionalism'. In *The Stanford Encyclopedia of Philosophy,* edited by E. N. Zalta.

Lewis, D. K. 1972. 'Psychophysical and theoretical identifications'. *Australasian Journal of Philosophy* 50:249–258.

———. 1983. 'Mad Pain and Martian Pain'. In *Philosophical Papers,* 1:122–132. Oxford University Press.

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

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

Paul, C. 2006. 'Morphological computation: A basis for the analysis of morphology and control requirements'. *Robotics and Autonomous Systems* 54:619–630.

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

Shoemaker, S. 1984. 'Some varieties of functionalism'. In *Identity, Cause and Mind,* 261–286. Cambridge: Cambridge University Press.