Forthcoming in *Cognitive Systems Research*

Extended cognition and the metaphysics of mind

Zoe Drayson

University of Bristol, UK

ABSTRACT

This paper explores the relationship between several ideas about the mind and cognition. The hypothesis of extended cognition claims that cognitive processes can and do extend outside the head, that elements of the world around us can actually become parts of our cognitive systems. It has recently been suggested that the hypothesis of extended cognition is entailed by one of the foremost philosophical positions on the nature of the mind: functionalism, the thesis that mental states are defined by their functional relations rather than by their physical constituents. Furthermore, it has been claimed that functionalism entails a version of extended cognition which is sufficiently radical as to be obviously false. I survey the debate and propose several ways of avoiding this conclusion, emphasizing the importance of distinguishing the hypothesis of extended cognition from the related notion of the extended mind.

KEYWORDS

extended cognition, extended mind, functionalism

1. Introduction

The hypothesis of extended cognition (hereafter HEC) is the claim that cognitive processes can and do extend outside the head. While cognitive scientists evaluate HEC by studying its explanatory role in empirical research, is not clear how ideas about extended cognitive processes relate to issues in the philosophy of mind, such as the nature of mental states and their relation to brain states. It has recently been claimed by Sprevak (2009), however, that the functionalist position in philosophy of mind entails HEC. Functionalism is a widely-held view which seems to allow us to acknowledge the physical underpinnings of mind without simply reducing mental states like belief or pain to their neural correlates: instead, mental states are defined by their functional relations rather than by their physical constituents. Sprevak argues that not only does functionalism entail HEC, but that the version of HEC entailed is so radical as to be obviously false. He concludes that there is something wrong with functionalism, and, if functionalism is the best argument for extended cognition, then HEC is unjustified.

In this paper, I introduce the standard argument for extended cognition (Clark and Chalmers 1998), and the principles and examples which are invoked to support it (section 2). I then review the basic tenets of functionalism and the intuitions behind it (section 3), before setting out Sprevak's (2009) claim that functionalism and extended cognition are more closely related than previously thought (section 4). I focus on Sprevak's argument that the version of extended cognition entailed by functionalism is so radical as to be false (section 5), and I propose two ways one might defend a more moderate version of

extended cognition (sections 6 and 7). I then turn to my attention to functionalism itself, and its role in the argument for extended cognition (section 8) before concluding with some thoughts about the relation between the idea of extended cognition and the notion of the 'extended mind' (sections 9 and 10).

2. Extended cognition

The claim that cognitive processes can and do extend outside the head was introduced by Clark and Chalmers (1998). They ask the reader to consider three scenarios, in which a person watches two-dimensional 'Tetris-like' geometrical shapes on a computer screen and answers questions about the potential fit of the shapes into the depicted 'sockets'. In each case, the person has different resources at their disposal. In the first scenario (1) the shapes do not move: the person must mentally rotate the shapes to assess their fit. In the second scenario (2), the person has the choice to either press a button to rotate the shape on screen (which is feasibly faster), or to mentally rotate the shape as before. In the third scenario (3), set 'sometime in the cyberpunk future', the person can rotate the shape on screen using a neural implant instead of pressing a button, or use standard mental rotation.

Scenario (1) clearly involves cognitive processes. Clark and Chalmers suggest that case (3) is on a par with case (1), such that we are inclined to see think of the neural implant in (3) as part of the person's cognitive system. They then point out that the cognitive processes involved in scenario (3) display the same computational structure as those

involved in scenario (2), where the person presses a button to perform the on-screen rotation. Clark and Chalmers argue that we should consider the second and third scenarios as cognitively equivalent, since the same sequence of tasks are being performed in both. Unless we're already committed to the *a priori* claim that cognition can only take place internal to the skin/skull boundary, why should we think that the external processes involved in (2) are not cognitive?

> "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 recognizing 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, 8)

The benefit of this approach is to offer cognitive scientists a different way of explaining performance on cognitive tasks. The act of rearranging one's Scrabble tiles in the tray to create words, for example, can be interpreted as part of the cognitive process, rather than as a complex succession of inputs to and outputs of cognition.

Once they have introduced the idea of extended cognitive processes and their explanatory possibilities, Clark and Chalmers consider whether similar arguments can be developed to show that the mind itself is extended:

> "Perhaps some *processing* takes place in the environment, but what of *mind*? Everything we have said so far is compatible with the view that truly mental states - experiences, beliefs, desires, emotions, and so on - are all determined by states of the brain. Perhaps what is truly mental is internal, after all?" (Clark and Chalmers 1998, 12)

(In highlighting this difference, Clark and Chalmers acknowledge that there is a distinction to be made between mere cognitive processing and genuinely mental phenomena; this is a distinction to which I shall return in Section 9.)

Clark and Chalmers put the Tetris-style argument to one side at this point and turn their attention to arguing that the mind itself might extend outside the head.

> "While some mental states, such as experiences, may be determined internally, there are other cases in which external factors make a significant contribution. In particular, we will argue that *beliefs* can be constituted partly by features of the environment, when those features play the right sort of role in driving cognitive processes. If so, the mind extends into the world." (Clark and Chalmers 1998, 12)

The well-known example introduced at this point involves the character of Otto and his notebook. Clark and Chalmers ask us to imagine Otto as an Alzheimer's patient who relies on environmental props to structure his life; meanwhile Inga retrieves her beliefs from her memory in a normal way. Inga and Otto are then compared in the following thought experiment:

> "Inga hears from a friend that there is an exhibition at the Museum of Modern Art, and decides to go to see it. She thinks for a moment and recalls that the museum is on 53rd Street, so she walks to 53rd Street and goes into the museum. It seems clear that Inga believes that the museum is on 53rd Street, and that she believed this even before she consulted her memory. It was not previously an *occurrent* belief, but then neither are most of our beliefs. The belief was sitting somewhere in memory, waiting to be accessed.

> Now consider Otto. Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. Today, Otto hears about the exhibition at the Museum of Modern Art, and decides to go see it. He consults the notebook, which says that the museum is on 53rd Street, so he walks to 53rd Street and goes into the museum." (Clark and Chalmers 1998, 12-13)

Clark and Chalmers argue that Otto's case and Inga's case are analogous, because the explanatory role played by Otto's notebook matches that played by Inga's memory: the content of Otto's notebook plays the same function as the content of Inga's dispositional belief. The mere fact that Inga's belief was inside her skull while Otto's was outside his skull does not seem to make any relevant difference to the explanatory role (or 'functional poise') of his mental state. The natural way to explain Otto's action is to say that he wants to go to the museum and *he believes the museum is on 53rd Street*, despite the fact that this belief is not inside his head. This commits us to the idea that at least some mental states can be external to the agent, and thus to the idea that the mind extends. As in the previous example, we can't simply state that mental states must be internal by definition.

> "To provide substantial resistance, an opponent has to show that Otto's and Inga's cases differ in some important and relevant respect. But in what deep respect are

> the cases different? To make the case *solely* on the grounds that information is in the head in one case but not in the other would be to beg the question." (Clark and Chalmers 1998, 14-15)

In the remainder of this paper, I'll be exploring the relationship between these arguments and functionalism in the philosophy of mind.

3. Functionalism

In the philosophy of mind, functionalism is the thesis that mental states are defined by their functional relations rather than by their physical constituents: what makes something a mental state of a particular type, such as a belief or a pain, depends on the way it relates to other states. A particular type of mental state, therefore, could be realized in multiple different ways. One of the motivations for adopting functionalism is that it allows us to make sense of what might be called the 'Martian intuition': the possibility that a creature which could have the same types of mental states as us despite major physical differences between them and us.

Most functionalists think that while any particular instance of a mental state in humans can be identified with the corresponding brain state, there is no one brain state which is necessary and sufficient for a creature to be in that general type of mental state. This allows that a creature with a different sort of brain from us could still have mental states like beliefs or pains, while allowing that each particular instance of a mental state they

had would be physically different from our own. One can imagine a Martian, for example, with a brain that was silicon-based rather than carbon-based like our own. According to the functionalist, this should not in itself pose a barrier to attributing the Martian with mental states. What matters is the way its silicon-based states relate to each other and to the relevant inputs and outputs: the way that its beliefs and desires interact to produce actions, for example, or the way that injury results in pain and the appropriate behavior. (It should be noted that not all forms of functionalism require preserving the Martian intuition; I shall discuss this further in Section 8.)

Clark and Chalmers (1998) do not explicitly present their position as a development of the functionalist program in philosophy of mind, making no mention of functionalism in their arguments for extended cognition or the extended mind. Several commentators, however, have claimed that the argument either for HEC or the extended mind – bearing in mind that the distinction between the two positions is rarely made – relies to a certain extent on broadly functionalist notions. Rupert (2004, 421) believes that the main argument for HEC "contains a clear functionalist strain", and Shapiro (2008, 7) argues that "support for extended minds often rests on the functionalist theory of minds". Similarly, Weiskopf remarks:

> "The argument for this thesis [HEC] rests on a simple, orthodox functionalist principle […] Unusual realizers are a staple of the functionalist literature. The hybrids described by advocates of extended minds differ only in lying outside of the normal brain-body system." (Weiskopf 2008, 266)

Most commentators, however, stop short of the claim that functionalism *entails* extended cognition or extended minds. They acknowledge that "functionalist theorizing alone does not resolve the issue of extended states" (Rupert 2004, 421), and that "the case for extended cognition is not going to follow *a priori* from a theory of mind" (Shapiro 2008, 14). Sprevak's claim goes further:

> "the relationship between functionalism and HEC goes beyond support for the relatively uncontroversial claim that it is logically or nomologically possible for cognitive processes to extend […] [F]unctionalism entails that cognitive processes do extend in the actual world." (Sprevak 2009, 503)

Sprevak argues that functionalism doesn't just support HEC or raise its possibility, but rather that functionalism entails that cognition is actually extended beyond the head.

4. The argument from functionalism to extended cognition

Sprevak (2009) asks us to consider the relation between functionalism and the Martian intuition. Sprevak claims that in order to preserve the intuition that such a creature could have the same types of mental state as us, functionalism has to be pitched at a relatively 'coarse-grain': various details have to be abstracted from or ignored. An overly fine-grained functionalism, Sprevak points out, would deny mentality to creatures who exhibited slightly different reaction times or different pain decay responses, for example.

Sprevak also proposes a 'fair treatment' principle, according to which we don't commit ourselves in advance to privileging internal over external components when considering cognitive systems – this is akin to Clark and Chalmers' idea, outlined in Section 2 above, that we shouldn't deny a process is cognitive merely on the grounds that is it outside the head.

Sprevak introduces the example of a Martian with an 'ink-mark' memory: his memory, rather than being stored in patterns of neural activity, is stored internally in a series of ink-marks. To store new information, the Martian activates a process to create new ink-marks. Accommodating the Martian intuition into one's functionalist approach dictates that these ink-marks count as dispositional beliefs, because they play the same role – at an appropriate grain parameter – as dispositional beliefs do in human minds. But once we accept that there are dispositional beliefs in this Martian case, and if we treat internal and external processes equivalently according to the fair-treatment principle, then there must also be dispositional beliefs in the case of Otto and his notebook from the Clark and Chalmers paper, according to Sprevak. Unless one is already committed to the claim that mental states must be inside the head, Sprevak claims, there is no difference between the ink marks inside the Martian's head and the ink marks inside Otto's notebook: if we are prepared to attribute beliefs in one case then we should be prepared to attribute them in the other case too.

> "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. […] But if the functional roles are set this coarse [as to allow for the Martian in question to have

> beliefs], then they are also satisfied by the Otto–notebook system. Therefore, Otto's notebook counts as an extended belief." (Sprevak 2009, 508)

Sprevak concludes that any brand of functionalism which preserves the Martian intuition will entail HEC. This is an argument which doesn't just establish the modal version of HEC – that extended cognition is possible (conceded by most of the opponents of extended cognition) – but that HEC is a fact about the actual world.

> "If functionalism is coarse-grained enough to admit possible intelligent Martians, then actual extended systems also qualify as mental. The claim is that one's attitude to Martian worlds commits one to the truth of HEC in the actual world." (Sprevak 2009, 512)

So far, this looks like a bonus to proponents of the extended cognition argument: Sprevak has shown that HEC – normally taken to be a controversial claim – is entailed by a reasonably widespread position in philosophy of mind. But Sprevak takes his argument further, and claims that the sort of extended cognition entailed by functionalist is one which even the proponents of HEC would reject. I will discuss this in the next section.

First, however, it is important to notice that Sprevak's version of the Martian intuition is slightly different from the one introduced in Section 3. As I see it, the key role of the Martian intuition in philosophy of mind is to highlight a problem that arises if we identify a psychological kind, e.g. pain, with a physical kind, e.g. the firing of C-fibres. This would entail that if we came across a creature, be it an octopus or a Martian, that seemed to display all the signs and responses indicative of pain but lacked C-fibres, we'd have to

conclude that it wasn't in pain after all. Functionalism, on the other hand, allows us to attribute to mental states to creatures who display the right functional profile, even if they are physically quite different. Sprevak is correct, therefore, to introduce the Martian intuition as the claim "that it is possible for creatures with mental states to exist even if such creatures have a different physical and biological makeup than ourselves" (Sprevak 2009, 507). However, Sprevak's version of Martian intuition doesn't just allow that psychological kinds can be realized in different *physical* ways (call this the 'physical Martian intuition'); it also allows that psychological kinds can be realized in different *psychological* ways (call this the 'psychological Martian intuition'):

> "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 (for example, between its memory and perception) may not match ours." (Sprevak 2009, 507-508)

This is why Sprevak claims that accommodating the Martian intuition is a matter of grain. Notice that accommodating the traditional *physical* Martian intuition is not a matter of how fine- or coarse-grained we set the functional roles.[1] A very fine-grained characterization of functionalism could allow for physically different creatures with mental states like ours, as long as those physically different creatures displayed the same

[1] I am grateful to Dan Stoljar for drawing this to my attention.

fine-grained psychological profile as us. Accommodating Sprevak's *psychological* Martian intuition requires that we adopt a liberal functionalism: that is, one which characterizes the functional profiles of mental states in broad and inclusive ways. (The term 'Martian intuition' will hereafter refer to Sprevak's psychological version unless stated otherwise.)

This distinction is important to bear in mind to appreciate Sprevak's ultimate conclusion that any brand of functionalism which preserves the Martian intuition is in trouble. Sprevak is *not* claiming that any functionalist who would consider attributing mental states to physically different creatures faces major problems. He is indicating that there are problems for the liberal functionalist who wants to ignore differences in fine-grained psychology. Problems for liberal functionalism are nothing new, however: Sprevak's main target is HEC, which he suggests is only feasible if it relies on liberal functionalism. This is a claim to which I'll return in Section 10.

5. Radical extended cognition and a *reductio* of functionalism

Sprevak argues that the version of HEC entailed by functionalism is so radical as to be evidently false. He claims that once one is prepared to accept the Martian intuition, it looks like we can conceive of all manner of internal Martian states as cognitive: a Martian could have a memory that worked like a library, or the internet, or an address book, and we'd still be prepared to think of this as an instance of memory. But once we

add the fair-treatment principle to this, it looks like we're committed to treating functionally similar cases alike with regard to their cognitive properties, whether they are internal or external. If we'd consider an *internal* process cognitive, we are committed to considering a functionally equivalent *external* process cognitive too. Sprevak's point is that for any imaginable (internal) Martian cognitive process, the functionally equivalent *extended* process is also cognitive: if a Martian whose memory worked like a library, the internet, or an address book would be considered a cognitive system, then when humans actually use libraries, the internet, or an address book, these extended systems should be considered extended *cognitive* systems. These seem to be more radical cases of extended cognition than Clark and Chalmers had in mind.

To push this point further, Sprevak introduces the example of a program for calculating the dates of the Mayan calendar, and asks us to think of it as an algorithm which could be run on a desktop computer or inside a Martian brain.

> "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 be unaware that it has this cognitive process." (Sprevak 2009, 517)

The idea here is that we can imagine a Martian who has an internal process which can calculate the dates of the Mayan calendar, but who has never used this process. Once we're willing to call this a *cognitive* process, then the fair-treatment principle compels us to count my never-used desktop computer program, which runs the same algorithm, as a cognitive process. Even the most radical proponents of HEC would want to deny this conclusion. Sprevak assumes that this version of HEC is obviously false, and so that it constitutes a *reductio* of the functionalist position from which it was derived.

6. A defence of moderate extended cognition (I)

Proponents of HEC do not, in general, want to claim that a computer program one has never used could be part of one's cognitive processes. Clark and Chalmers distinguish acceptable cases of extended cognition, such as Otto's notebook, from unacceptably radical cases. They claim that the case of Otto and his notebook has features which aren't exhibited by cases involving (for example) never-used programs on one's computer:

> "First, the notebook is a constant in Otto's life – in cases where the information in the notebook would be relevant, he will rarely take action without consulting it. Second, the information in the notebook is directly available without difficulty.

> Third, upon retrieving information from the notebook he automatically endorses it" (Clark and Chalmers 1998, 17)

One might think that the same sort of distinction between acceptable and overly-radical applications of HEC can be used to counter the *reductio* considered in Section 5, but Sprevak denies this is the case. Sprevak alleges that Clark and Chalmers' attempt to distinguish between the Otto case and overly-radical cases such as the Mayan calendar example must be rejected, on the grounds that it violates their own position on the fair treatment of internal and external processes. Recall that Clark and Chalmers suggested that if a partly external process would be considered cognitive were it done in the head, then we should consider that process cognitive. According to Sprevak's version of this claim (the fair-treatment principle), the principle should work both ways: we can't set standards for potential *extended* cognitive processes that aren't met by existing *internal* cognitive processes. Sprevak claims that there are *internal* human examples of cognition which wouldn't be considered cognitive on Clark and Chalmers' above conditions of typicality, availability, and endorsement. He points out that there are resources used in acts of outstanding human creativity which aren't typically invoked, that the information in my visual system about the position of my eyes isn't directly available to me, and that we don't automatically endorse the outputs of our imagining or desiring processes. Furthermore, Sprevak argues, Clark and Chalmers can't add or substitute different conditions: due to the liberalness of the Martian intuition, *any* conditions added to an argument for extended cognition to restrict it to a moderate version will be violated by cases of hypothetical, if not actual, internal cognition. And if extra conditions for

extended cognition are added which do not have to be (and indeed are not) met by cases of internal cognition, Sprevak argues, then this goes against the idea of equal treatment for internal and external processes.

It strikes me, however, that Clark and Chalmers' position is subtly different from Sprevak's portrayal. He claims that Clark and Chalmers introduce the notions of typicality, availability, and endorsement as "extra conditions to the functionalist credo" (Sprevak 2009, 514) which are individually necessary conditions for cognition. I maintain that Clark and Chalmers are not making any claims here about necessary or sufficient conditions for cognition. When Clark and Chalmers mention that the information in Otto's notebook is directly available, typically invoked, and automatically endorsed, they are instead highlighting ways in which the contents of Otto's notebook have the same functional poise as our own (and Inga's) normal dispositional beliefs. They are not using these features of the notebook's contents to characterize what makes something *cognitive* rather than *non-cognitive*, but merely claiming that typicality, availability, and endorsement are features of *dispositional beliefs*.

Given that the criteria used by Clark and Chalmers to support the function of Otto's notebook in his mental life are only concerned with the specific case of dispositional belief, one could argue that it is simply irrelevant whether there are other forms of cognition (e.g. those involved in human creativity, the visual system, or desire and imagination) which are not typically invoked, readily available, or automatically endorsed. But there is a worry that the same argument can be run for dispositional beliefs:

are there internal cases of dispositional belief that don't meet the criteria of availability, typicality, and endorsement? If we'd concede, for example, that someone could have a dispositional belief that was only available to them after a good night's sleep, then it looks like we can't use the criterion of availability as a way to differentiate between moderate cases of extended belief like Otto's and more radical cases.[2]

When we start to explore what would and would not count as a dispositional belief, we start having to rely on our intuitions. Take the Martian with the never-used Mayan calendar program in his head, for example. Does he have beliefs about Mayan dates? Would we say he remembers them? Probably not. We might be prepared to acknowledge that the process in the Martian's head is a cognitive process, as outlined in Section 5, but when we're pushed to say what *sort* of cognitive process, it becomes more difficult. If we can't specify what the cognitive relationship is between the Martian and the program, why are we so sure there is one? I shall explore this idea further in the next section, but the important point to note here is that once we shift the debate to specific cognitive processes rather than cognition in general, our intuitions can change.

7. A defence of moderate extended cognition (II)

[2] This assumes that we accept the fair-treatment principle of treating internal and external cases equally, but given that we need it to work in one direction to allow for extended cognition in the first place, it's not clear how we could refuse to apply the same principle in the other direction. One might also argue that Clark and Chalmers' criteria are not intended to be necessary and sufficient conditions even for dispositional belief, but then the project of drawing a distinction between moderate and radical cases of extended cognition becomes even harder.

Another way to block Sprevak's *reductio* would be to challenge his argument about the Mayan calendar algorithm. Wheeler (2010) does just this, by claiming that Sprevak's argument relies on a stronger version of the Martian intuition than that generally assumed by functionalism. Wheeler reconstructs Sprevak's Mayan calendar argument (outlined in Section 5 above) roughly as follows:

i) Take a non-cognitive, externally-located element X

ii) Imagine X inside the head of a Martian

iii) Suggest that the Martian system is cognitive (using the Martian intuition)

iv) Infer that X also deserves cognitive status (using the fair-treatment principle)

v) Derive a contradiction from (i) and (iv)

vi) Acknowledge that functionalism entails HEC (from the argument in Section 4)

vii) Use the contradiction in (v) as a *reductio* of functionalism

Step (iii) of this argument is the critical one. Wheeler suggests that Sprevak has used the Martian intuition to argue that a non-cognitive, externally-located program acquires cognitive status when placed inside the head of a Martian. If this is the case, it looks to require a stronger version of the Martian intuition than previously considered. Traditionally, the Martian intuition is the idea that a creature could possess cognition like us despite being physically very different; Sprevak's use of the Martian intuition in step (iii), on the other hand, seems to require that an element can be considered cognitive purely by virtue of being located inside a creature's head. If Sprevak's argument can be

blocked at step (iii), then the *reductio* of functionalism fails, and the possibility that one could hold a moderate version of HEC is left open.

It should be noted that Sprevak isn't making the radical claim that we can use the Martian intuition to judge *any* internal Martian element to have cognitive status. We presumably want to allow that elements of the brain governing activities such as respiration and digestion are not cognitive, not to mention the intuition that foreign bodies introduced into brains do not thereby become part of the cognitive system. (Were the latter the case, Phineas Gage could replace Otto as the mascot of the extended cognition movement.) Sprevak's point is that the element in question here is a program for calculating dates. This seems to be precisely the sort of information-processing algorithm which, *hooked up in the right way* inside the head of a cognitive agent, would be considered cognitive. This is what Wheeler would have to deny in order to block the *reductio*.

As Wheeler goes on to point out, a lot hinges here on how we cash out this idea of being hooked up 'in the right way'. Whatever functionality we need to assume to allow the internal program cognitive status at step (iii) of the argument must then be transferred to the external program at step (iv) of the argument, in accordance with the fair-treatment principle. And the more we focus on the integration between the program and its surroundings, the more doubts we can introduce regarding the cognitive status of the Martian element, or indeed the non-cognitive status of the external element. Our intuitions, as Wheeler points out, can also be shifted by subtle changes to the hypothetical scenario, such as the order in which we consider the Martian case and the external case.

> "When we begin our reflections on the issues, as Sprevak does, by focusing on an example of a desktop software program, our natural tendency is to think of an isolated and easily removable software application […] On the other hand, when we begin our consideration of the issues by imagining the Martian inner program, our natural tendency is to think of that mechanism as being already functionally integrated into (although not yet activated within) an organized economy of states and processes. […] If the desktop program for calculating the Mayan calendar were a functionally integrated element in this kind of economy, then it may seem far less crazy to conclude that it could be a cognitive mechanism, or at least, part of one, even though it is spatially located outside the head." (Wheeler 2010, 267-268, note 8)

Sprevak's *reductio* of functionalism, considered this way, relies more heavily than it might have seemed on our intuitions about when to consider two processes functionally similar, and when to consider a process cognitive. I'll return to the issue of our intuitions regarding minds and cognition in Section 9, but in the meantime it's worth examining how closely related functionalism and HEC actually are.

8. Some thoughts about functionalism

In Sections 6 and 7 I focused on the second part of Sprevak's argument: the attempt to show that the form of extended cognition entailed by functionalism is so radical as to be obviously false, thus providing a *reductio* of functionalism. I argued that it might be possible to defend a moderate version of extended cognition, perhaps by getting clearer

on how to understand Clark and Chalmers' (1998) 'conditions' of typicality, availability, and endorsement, or by paying closer attention to how our intuitions about 'Martian' cognition and the fair treatment of internal and external processes.

Another way to avoid Sprevak's conclusion would be to take issue with the first part of Sprevak's argument (outlined in Section 4 above), regarding the initial claim that functionalism entails HEC. Sprevak argues that any version of functionalism which is sufficiently coarse-grained as to accommodate the Martian intuition will entail HEC. One might wonder why functionalism needs to accommodate the Martian intuition in the first place: must a functionalist be committed to the metaphysical task of giving a solution to the mind-body problem? In fact, not all varieties of functionalism are interested in attributions of mentality to non-human hypothetical creatures (as Sprevak himself acknowledges). Psychofunctionalism, for example, might aim merely to capture generalizations relevant to theories of human psychology. Psychofunctionalists are committed to the idea that mental states and processes should be introduced and individuated in terms of their roles in producing behaviour (i.e. functionally), but according to the best scientific theories of behavior.

The problem for this psychofunctionalist is that it looks like arguments for extended cognition won't even get off the ground on this characterization of functionalism. If our best scientific theories tell us that cognition is a function of brains and that cognitive processes are internal processes, then HEC is a non-starter. This is not the claim that extended cognition is impossible or incomprehensible, rather that it runs counter to what

we know about cognition. HEC requires Otto's relation to his notebook to be sufficiently similar to Inga's relation to her memory for the two cases to qualify as instances of the same sort cognitive processes. But the functional profile of normal memory involves, for example, primacy and recency effects which aren't a feature of the Otto's notebook.

> "although in principle, cognitive science could serve up extension-friendly psychofunctionalist roles, thus far it has not. The most interesting and useful profiles of psychological states and properties detail causal roles that external materials are not likely to fill. […] The empirical study of memory, for instance, reveals many fine-grained properties that external resources—notebooks, for example—do not, for the most part, exhibit." (Rupert 2009, 95-96)

One might try to argue that these 'fine-grained properties' of normal cognitive processes are irrelevant, because we can imagine all sorts of subtly different ways that a creature's memory might work and still fall under our general concept of memory. But this looks to put us back where we started, because it requires hypothetical attributions of mentality to non-actual minds. This psychofunctionalist, by assumption, is only interested in capturing the generalizations of scientific psychology. This seems to back up Sprevak's argument that fine-grained functionalism won't give us extended cognition, but once we opt for a coarser grained functionalism, we open the door to Martians and radically extended cognition.

> "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). […] 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" (Sprevak 2009, 510)

Sprevak's case seems to be a solid one: proponents of HEC who rely on liberal functionalism to make their case will face the difficult task of restricting extended cognition to moderate cases (see Sections 6 and 7), while if they constrain the functional roles to avoid this conclusion, they face the problem that their own moderate cases of extended cognition don't count as cognition any more.

Can the proponent of HEC make their case without relying on functionalism? Sprevak himself thinks that "[t]he most plausible justification of HEC is the functionalist argument" (Sprevak 2009, 527) and, as we saw in Section 3 above, the idea that HEC is in some sense reliant on broadly functionalist notions "is now pretty much part of the received view of things" (Wheeler 2010, 245). The commitment to functionalism is usually located in what has become known as the 'parity principle', the statement in Clark and Chalmers (1998) on which Sprevak's fair-treatment principle draws:[3]

> "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 recognizing 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, 8, italics in original)

Rupert, although he doesn't think the functionalist argument for HEC works, acknowledges that even Clark and Chalmers seem to think that the parity principle "is

[3] It should be noted at this point that Sprevak (2009) explicitly separates the 'parity principle' and functionalism: his argument is that HEC and associated problems only arise when the two are combined.

meant to direct our attention to what are the foundational considerations: a consideration of functional role" (Rupert 2009, 89):

> "There is a straightforward connection between functionalism and the Parity Principle: The *location* of a state does not matter, according to the functionalist view, so long as the state in question plays the appropriate causal–functional role […] and on one way of reading the Parity Principle, it directs us to ignore location when evaluating something's cognitive status." (Rupert 2009, 89-90)

Chalmers, however, has recently suggested that the relationship between functionalism and HEC is weaker than generally thought.

> "This [the claim that HEC relies on functionalism] cannot be quite right: I think that functionalism about consciousness is implausible, for example, but this implausibility does not affect the arguments for the thesis. One might support the view by invoking an attenuated functionalism: say, one where certain mental states (such as dispositional beliefs) are defined by their causal relations to conscious states, to behavior, and to other elements of the cognitive network. I find such a picture attractive myself, but strictly speaking even this picture is not required for all the argument to go through." (Chalmers 2008, xv)

All one needs, claims Chalmers, is "the very weak functionalism" captured by the idea that "if a state plays the same causal role in the cognitive network as a mental state, then there is a presumption of mentality" (Chalmers 2008, xv). In the following sections, I will develop Chalmers' thoughts on this point, and argue that the 'parity principle' need not even involve weak functionalism. To do so, however, I need to return to the distinction between extended cognition and the extended mind.

9. Mind and cognition

When I outlined the structure of Clark and Chalmers' arguments for HEC in Section 2, I was careful to distinguish between their argument for extended *cognitive processes* and their argument for the extended *mind*; a distinction has been largely overlooked in the ensuing literature (although see Rupert [2004] for a brief but illuminating discussion). The distinction between extended cognition and extended minds is presumably derivative on the distinction between cognition and the mind. In this section, I'll look at the relationship between the sciences of cognition and the philosopher's concept of the mind, with a view to clarifying the debate over extended cognition and extended mind in the following section.

What is mind? *Mental* states are traditionally thought of as the ascriptions of 'folk-psychology': the commonsense everyday states we use to describe, explain, and predict our fellow human beings' actions. Prime examples of mental states are experiential states (such as pains and other sensations), and thoughts: propositional attitudes such as beliefs, desires, hopes, fears, expectations, and intentions. The relations between thoughts are rational: whether my belief and my desire produce a corresponding intention depends on what my belief is *about* and what my desire is *for*. And to have thoughts, it is widely believed that one must have the appropriate *concepts* which can combine and recombine

to produce new thoughts. Thoughts are generally assumed to be the sorts of states of which one is conscious, or at least potentially so.

Several more general things can also be said about minds: the concept of mind is entwined with both our everyday and our philosophical concepts of selfhood and persons, of rationality and responsibility, of free-will and morality. We tend to assume that only adult humans have minds in this sense, but even then there is disagreement about what the conditions for 'mindedness' are. Philosophers have variously suggested that only creatures having *conscious* thoughts can count as having thoughts at all (Searle 2002), that to have a mind one must be an interpreter of the speech and behavior of another (Davidson 1975), and that genuinely mental agents must be capable of appreciating the normative force of a reason for belief or for action (McDowell 1994).

What is cognition? The *cognitive sciences* – including psychology, artificial intelligence, linguistics, robotics, and neuroscience – aim to understand minds and intelligent behavior by studying its mechanisms (where these are usually understood to be computational). When cognitive science tries to explain traditional mental phenomena, it can end up positing different processes to those of our everyday understanding. Our commonsense notion of 'remembering', for example, is broken down by the cognitive sciences into procedural memory and declarative memory, and the latter is further broken down into episodic memory and semantic memory.

In addition to categorizing the domain of the mental differently, cognitive science also *expands* upon the traditional domain of the mental. It is not just interested in the consciously accessible states we can attribute to a person, but also the inaccessible states we can only attribute to parts of a person's cognitive system. Chomsky, for example, argued that much of our 'knowledge' of basic grammar isn't actually something we can really be said to know, or to believe, or to think; he suggested that the term 'cognizing' was more appropriate:

> "Thus 'cognizing' is tacit or implicit knowledge […] cognizing has the structure and character of knowledge, but may be and in the interesting cases is inaccessible to consciousness." (Chomsky 1980, 69-70)

This information about grammar is available only to certain parts of a person's language system: it is not information that a person can actually use in their general reasoning processes. It is both inaccessible to conscious experience and informationally isolated from thought. The workings of the early visual system show a similar profile to that of basic grammar. In vision, various processes use aspects of contrast and light to build up the scene that one can eventually perceive and extract information from. But in the early stages, these processes are not consciously accessible and the information they contain (about e.g. zero-crossings) is not available to the person, but only to the next part of the visual process. Given that the information in both the grammar case and the visual case is isolated from a person's general background knowledge, it cannot be used in inferences or combined with a person's thoughts: the information is not conceptual in the way that the constituents of thoughts are. A person might need to possess the concept of a zero-crossing in order to have beliefs about zero-crossings, but they arguably don't need this

concept to have a working visual system which contains information about zero-crossings.

The term ‘cognition’, therefore, has come to be used to refer to all the mechanisms responsible for mental phenomena; cognition encompasses traditional mental phenomena – although possibly in a non-traditional way – and also the lower-level information processes which account for them. Although Clark and Chalmers (1998) do not elaborate upon their distinction between cognition and mind, Clark (2008) makes the difference clear:

> “The term 'cognizing' is here used to mark a notion of the mental that is broader than the one suggested by introspection and common-sense alone. Where introspection and common-sense might identify mind simply as a locus of beliefs, desires, hopes, fears etc, the scope of the cognitive may include states and operations unearthed by science. Examples might include grammars (if psychologically real), and the states and operations implemented by low-level vision.” (Clark 2008, footnote 4, introduction)

Given these distinctions, there seem to be things we can say about mental states and processes that we can’t say about cognitive states and processes, and vice versa. It remains to be seen whether the states picked out by folk psychology, such as beliefs and desires, map neatly onto the states and processes picked out by cognitive science. If there is no neat correspondence, then there are various ways that mind and cognition might relate. One might argue that if traditional mental states elude analysis in terms of the

states which cognitive science eventually endorses, then cognitive science has simply failed in its remit to investigate the nature of mind: whatever discoveries cognitive science has made, they are not discoveries about the mind. This is a view associated with philosophers who think that we can learn everything there is to know about the mind by *a priori* reflection on and analysis of our concepts of mentality:

> "such philosophers' methodological commitments assign no special authority to science. On their view, we discover the nature of the self, the mind, and cognition, by reflecting on our concepts." (Rupert 2009, 11)

Alternatively, one might take commonsense to be at fault, and suggest that our everyday picture of the mind is wrong and should be corrected by science.

> "[Those] who take completed cognitive psychology to issue the final word about human cognition might take the arguments [of cognitive science] to bear strongly on questions about the ultimate nature of the human mind and self" (Rupert 2009, 11)

In the extreme case, this might mean losing such terms as 'belief' and 'desire' from explanations of intelligent action. Another option would be to acknowledge that much of our 'mental state' talk in folk psychology is not scientific, but nonetheless still philosophically relevant to the understanding of mind (Clark 1989, 46). A lot depends on the outcomes of future empirical research and philosophical responses to it, but it seems fair to say that, at present, "the place of mind in cognitive science is highly problematic" (Clark 1989, 1).

10. Extended cognition and extended minds

Now I return to *extended* cognition and the *extended* mind: it looks like how one understands the relation between them will depend on how one understands the relation of mind to cognition. In particular, an argument for extended cognition will not, in most cases, also provide an argument for the extended mind.

First, the extended cognitive processes in question might be outside the traditional domain of the mental: they might be processes which are not attributable to a person, but only to a part of that person's cognitive system, and involve non-conceptual information which is neither integrated with their thoughts nor accessible to their consciousness. Such processes might prove to be a causal precursor to genuinely mentality, but this wouldn't be sufficient to show that the mind itself extends. Second, even if the cognitive processes in question are responsible for traditionally mental phenomena, it is plausible that the categories of cognitive science may cross-cut those of our commonsense idea of the mind, such that it is only the internal part of a particular cognitive process which is relevant to that particular mental state. Third, it is open to the philosopher to simply deny that the findings of cognitive science can have any implications for the mind.

Despite these possibilities, among the few who make the distinction between extended cognition and the extended mind, it is generally considered that "current work on extended cognition promises to provide the strongest support to date for the view that the mind is extended" (Rupert 2004). This argument structure can be found in the Clark and

Chalmers paper: they argue for extended cognition using the Tetris example, and then use this to develop an argument for the extended mind, supported by the idea of Otto and his notebook.

How does this distinction between mind and cognition relate to Sprevak's argument? Sprevak's argument concerns the claim "that *cognitive* processes can and do extend outside the head. Call this the 'hypothesis of extended *cognition*' (HEC)" (Sprevak 2009, 503, my italics). But he introduces both thought experiments from the Clark and Chalmers paper, the Tetris case and Otto, as part of the argument for HEC (Sprevak 2009, 504), which suggests he does not distinguish between extended cognition and the extended mind. Furthermore, Sprevak sometimes switches from talking about cognition to question, for example, whether "one could appeal to general theories of *mentality* to decide whether extended cases were *mental* or not" (Sprevak 2009, 523, my italics). In failing to distinguish between extended cognition and the extended mind, Sprevak is in good company: most commentators on the debate over the past decade have done similarly. In many cases, it makes little difference to the discussion. In Sprevak's case, however, we are concerned with the relationship between these claims and functionalism; and here, I maintain, the distinction matters.

One of Sprevak's main concerns in his paper is to demonstrate that the "reasons for HEC's failure bring to light new troubles with functionalism as an account of cognitive systems" (Sprevak 2009, 503). I want to suggest that we take care to make another distinction, namely that between functionalist accounts of the mind and computationalist

accounts of cognition. Functionalism is the metaphysical view that mental states are individuated by their functional relations with mental inputs, outputs, and other mental states; computationalism is the hypothesis that the functional relations between cognitive inputs, outputs, and internal states are computational. Computationalism is neutral on whether the computational relations it posits constitute the nature of mental states, and functionalism is neutral with regard to the characterization of the functional relations it posits. The two positions, however, are often conflated: when Fodor, for example, points to "the widespread failure to distinguish the computational program in psychology from the functionalist program in metaphysics" (Fodor 2000, p.105, note 4), he singles out an earlier work of his own as a prime example.[4]

With this distinction in mind, I'd like to return to the widespread claim (considered in Section 8) that the 'parity principle' relies on functionalism. This principle requires that the same cognitive processes can be realized in more than one way, and functionalism is an obvious way to accommodate multiple realizability, so it is natural to interpret the 'parity principle' as a functionalist assumption. Multiple realizability, however, is also exhibited by computational algorithms: the same program can be run by different physical systems. I suggest that the 'parity principle' can be understood as a claim about computational accounts of cognition.

[4] For a detailed discussion of the differences between computationalism and functionalism, and the historical reasons for their conflation, see Piccinini (2004).

Going back to the Tetris example, recall that Clark and Chalmers wanted to argue that the external button-pressing process in the second scenario should be considered equivalent to the futuristic process involving the neural implement in the third scenario:

> "And case (2) with the rotation button displays the same sort of *computational structure* as case (3), although it is distributed across agent and computer instead of internalized within the agent. If the rotation in case (3) is cognitive, by what right do we count case (2) as fundamentally different?" (Clark and Chalmers 1998, 8, my italics)

I take this to a claim about computational equivalence: two computational systems are computationally equivalent when they compute the same function in the same way: in more technical terms, when they are input-output equivalent and compute the same algorithm. Clark and Chalmers' suggestion that an external process should be considered cognitive when the equivalent internal process would be so considered can be understood as expressing a commitment to computational accounts of cognition without any commitment to functionalism.

So can we argue for extended cognition without functionalism? I've argued that we can understand the equivalence between the third Tetris case with the neural implant and the second Tetris case with the button-pressing as the computational equivalence of cognitive processes, which doesn't require that we be functionalists about mental states. But now we're left with the claim that if the neural implant process is cognitive, then so is the button-pressing process: why should we think the processing involved in the implant case is *cognitive* processing? There's a worry that the obvious way to argue that a hypothetical

case is a case of cognition will involve some sort of Martian intuition.[5] But it seems that where cognition (rather than mentality) is concerned, we'd be a lot happier simply coming up with some criterion to distinguish between the cognitive and the non-cognitive,[6] or just defining cognition ostensively (and in such a way as to avoid Rupert's worry). But ultimately, this makes cognition a lot less philosophically interesting than minds and mentality: to assert that cognitive processes extend beyond the head is not to make a claim about the *nature* of cognition.

The claim that the *mind* extends is the more philosophically interesting hypothesis, and I suggest that there are two ways to argue for it. One can argue for it directly using the Otto thought experiment, but this requires functionalism about the mental. And it looks like Sprevak is right on this one: any brand of functionalism which is liberal enough to get the argument going in the first place will end up positing radically extended minds.

The other way to argue for the extended mind is to show that cognition is extended, while claiming that the relationship between mind and cognition is one which entails that if cognition is extended, then mind is extended. One way to get such a relation is by stipulation: by claiming that minds are simply whatever science determines them to be. In this case it would be trivial that minds extend if cognition extends, but this would have few interesting philosophical implications for idea about the self or personhood, for example. A different option would be to retain a distinct notion of mind as a collection of properties which 'emerge' from cognition, but in this case it's not clear what it would

[5] Thanks to Mark Sprevak for emphasizing this point in personal communication.
[6] See Wheeler (2010) for some ideas about how to develop a 'locationally-uncommitted' mark of the cognitive.

mean for the mind to be extended. In particular, it would be difficult to say of any particular mental state that it was extended, because the categorization of mental states needn't map neatly onto the cognitive processes. A third possibility would be for the causal structure of our cognitive systems to precisely mirror the posits and relations of our everyday mental ascriptions: in other words, for there to be a 'language of thought'. In this case, if we could show that a particular cognitive structure extends beyond the head, then we could infer that the corresponding mental state was extended. But to argue for the extended mind on such a basis is to play hostage to empirical fortune.

Acknowledgement

Earlier versions of this paper were presented at the University of Edinburgh, the University of Bristol, and the Australian National University. I have benefited greatly from discussions with Ken Aizawa, Dave Chalmers, Andy Clark, Rob Rupert, Larry Shapiro, Sven Walter, and Dan Weiskopf. I owe a particular debt of thanks to Mark Sprevak and Mike Wheeler who both commented extensively on an earlier draft.

References

Chalmers, D. (2008) Foreword. In A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.*. New York: OUP.

Clark, A. (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.*. New York: OUP.

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19.

Fodor, J.A. (2000), *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*. Cambridge, Mass.: The MIT Press.

Piccinini, G. (2004). Functionalism, computationalism, and mental states. *Studies in the History and Philosophy of Science*. *35*, 811–833.

Rupert, R. D. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy, 101*(8), 389-428.

Rupert, R. D. (2009). *Cognitive Systems and the Extended Mind*. New York: OUP

Shapiro, L. (2008). Functionalism and the boundaries of the mind. *Cognitive Systems Research*. *9*: 5-14.

Sprevak, M. (2009). Extended cognition and functionalism. *Journal of Philosophy.*

Weiskopf, D. A. (2008). Patrolling the mind's boundaries. *Erkenntnis, 68*(2), 265-276.

Wheeler, M. (2010), In defense of extended functionalism. In R. Menary (ed.) *The Extended Mind* (245-270). Cambridge, Mass.: MIT Press.