Functionalism and the Case for Modest Cognitive Extension

Mikio Akagi

Examination No. 9772703
Taught MSc in Philosophy
The University of Edinburgh
Submitted 21 August, 2009

Revised 11 November, 2009

Abstract: The Hypothesis of Extended Cognition (HEC) holds that that not all human cognition is realized inside the head. The related but distinct Hypothesis of Extended Mentality (HEM) holds that not all human mental items are realized inside the head. Clark & Chalmers distinguish between these hypotheses in their original treatment of cognitive extension, yet these two claims are often confused. I distinguish between functionalist theories on which functional roles are individuated according to computational criteria, and those on which functional roles are individuated according to rational criteria. I then present an argument for a modest version of HEC from computational functionalism, based on Clark & Chalmers' original argument. In doing so I articulate a successor to their parity principle, and review studies by Wayne Gray et al. that provide plausible evidence for actual cognitive extension. I then respond to a new criticism of HEC by Mark Sprevak using the modest account I have developed, arguing that Sprevak conflates HEC and HEM.

I. Introduction[1]

Andy Clark and David Chalmers have argued, infamously, that not all cognition happens inside the head. This claim, which has become known as the *Hypothesis of Extended Cognition* (*HEC*),[2] has attracted many critics. Despite numerous re-treatments of the hypothesis of extended cognition by Clark and other authors, however, misunderstandings of cognitive extension are widespread. Even smart, recent criticisms by Fred Adams & Ken Aizawa, Robert Rupert, and Mark Sprevak betray misinterpretations of the consequences of cognitive extension, and of the dialectic setting of its arguments. The best remedy for this confusion may be to go back to the basics. In this paper, I will present an argument for the modest core of HEC, based on Clark & Chalmers' original argument and some of Clark's more recent remarks. I will be concerned to relate Clark & Chalmers' arguments to the structure of functionalist theories, and in particular to be careful about the distinction between functionalist theories that are based on a methodology of computational individuation, and those that are based on a methodology of rational individuation. Careful attention to the distinction between computation and rationality will, I contend, protect HEC from certain forms of objection, of which I will take Sprevak's as an articulate example.

Cognitive Extension, Take 1

Before introducing Sprevak's criticism, I will briefly review the highlights of Clark & Chalmers' original discussion. The centrepiece of Clark & Chalmers' argument has come to be known as the *parity principle*:

> 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 a cognitive process, then that part of the world *is* (so we claim) part of the cognitive process.[3]

---

[1] This paper has benefited significantly from the comments and counsel of my advisor Jesper Kallestrup, and from extensive and discussions with Lily Prudhomme Copple, Evan Butts and Amber North. I am also thankful for the direction I received from Andy Clark and David Levy, and the significant influence on my thoughts about cognition and mentality that was wrought by formative lessons from Alan Baker and Richard Eldridge. The faults remaining in the paper are, of course, mine.

[2] This is the term is used in Rupert (2004, forthcoming-a), Clark (2008), and Sprevak (forthcoming).

[3] Clark & Chalmers 8. The parenthetical qualification is sometimes quoted 'for that time' instead of 'so we claim,' e.g. in Clark 2008 p. 77 (but not in the appendix on p. 222). This

Clark & Chalmers try out the principle in two thought experiments, each of which is meant to support a different claim. The first claim is the hypothesis of extended cognition, HEC. This is the claim with which I am primarily concerned in this paper. However, Clark & Chalmers also argue for a second claim, that the *mind* is extended. I will call this claim the *Hypothesis of Extended Mentality*, or *HEM*. Their second famous thought experiment about Inga and Otto is deployed in service of establishing this second claim.

In their first thought experiment, Clark & Chalmers ask us to imagine the following three scenarios: (T1): A person tries to determine whether a geometric shape displayed on a monitor fits into a slot by mentally rotating it. (T2): A person tries to determine whether a geometric shape fits into a slot, but can either rotate the shape mentally or press a button to have a computer perform the rotation and display the rotated shape on the screen. We can assume that the latter option is typically faster than the former. (T3): A person tries to determine whether a geometric shape fits into a slot; the person can rotate the shape mentally, or activate a neural implant that can perform the rotation just as the computer can in (T2) and display it.[4]

Clark & Chalmers argue that if both the options in (T3) count as cognitive processes, then the parity principle entails that both options in (T2) so count. The cyberpunk implant has *ex hypothesi* identical fine computational structure as the computer program, and could even be activated by motor processes and feed the results of its rotation into the person's visual system. Finally, if both options in (T2) count, then cognition is in fact extended, since (T2) describes one component of playing the video game Tetris. Clark & Chalmers cite research by David Kirsh and Paul Maglio[5] that experienced Tetris players supplement their slow biological resources for mental rotation by using the computer to rotate pieces more quickly than they can with native resources. Clark & Chalmers then go on suggest that there are many other circumstances where people (often unconsciously) exploit their environment, using external processes in lieu of head-internal cognitive processes. Such cases, they claim, are cases of extended cognition.

This comparison between (T2) and (T3) certainly seems apt, but the case for HEC still depends on arguing that the operation of the cyberpunk implant can be a part of a cognitive system. (T3) is a handy intuition pump, but it is

---

alternate qualification is not strictly necessary, but is illuminating in its own way—particularly against some glib caricatures of Clark & Chalmers, as in Adams & Aizawa (forthcoming), qtd in Clark 2008, p. 86.

[4] Clark & Chalmers 7

[5] 1994, cited in Clark & Chalmers

hardly an uncontroversial case of cognition. That the implant is inside the head should not be sufficient for it to realise cognitive processes—a stone would not realise cognitive processes simply because it is inside someone's head. It is this early stage in Clark & Chalmers' argument that some objections to HEC insinuate themselves.

After their discussion of HEC, Clark & Chalmers acknowledge that their second claim, HEM, has not been established:

> 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.[6]

In order to establish HEM, Clark & Chalmers propose a second thought experiment. Imagine that one day Inga decides that she would like to see an exhibition at the Museum of Modern Art in New York, where she lives. She recalls that the museum is on 53rd Street, so she walks to 53rd Street and views the exhibition. Otto also decides that he would like to see the exhibition at MoMA, but since Otto suffers from Alzheimer's disease he uses extra-bodily artefacts to compensate for the failings of his biological memory. In fact, Otto has a notebook that he carries with him everywhere, in which he writes information he learns and which he checks frequently in the course of his daily activities. When Otto hears about the exhibition at MoMA he decides that he would like to see it. He checks his notebook, sees that MoMA is on 53rd Street, and sets off for the museum. Clark & Chalmers propose that

> Otto walked to 53rd Street because he wanted to go to the museum and he believed the museum was on 53rd Street. And just as Inga had her belief even before she consulted her memory, it seems reasonable to say that Otto believed the museum was on 53rd Street even before consulting his notebook… The information in the notebook functions just like the information constituting an ordinary non-occurrent belief; it just happens that his information lies beyond the skin.[7]

Although Otto may not exist, Clark & Chalmers argue that it is not out of the question that an external device like a notebook could come for some person to realise some of the roles associated with beliefs.

Clark & Chalmers express uncertainty about how liberally instances of HEM should be identified. They suggest several scenarios besides Otto that

---

[6] Clark & Chalmers 12

[7] Clark & Chalmers 13

seem roughly analogous, including the use of a filofax and the realisation of beliefs and desires in the mental states or linguistic performances of other people.[8] However, whereas Clark & Chalmers claim that *if* such a person as Otto existed then he would have extended beliefs, they refuse to commit to whether these other cases would be instances of HEM. They say merely that 'we do not think that there are categorical answers to all of these questions, and we will not give them."[9] Instead, they suggest four apparently *ad hoc* conditions that make the Otto example seem relatively acceptable:

> First (H1), 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 (H2), the information in the notebook is directly available without difficulty. Third (H3), upon retrieving information from the notebook he automatically endorses it. Fourth (H4), the information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequences of this endorsement.[10]

Clark & Chalmers express reservations about these conditions, however, and in particular are sceptical about the fourth.

This entire discussion is very puzzling. Clark & Chalmers suggest that they have an argument for HEM and endorse one instance of it, but back off from other possible instances of HEM citing obviously *ad hoc* conditions and waving their hands like mad. But before I explain what I think is happening here, I will discuss an objection to Clark & Chalmers' argument that intercedes at this juncture.

Sprevak's Challenge

My first pass at Clark & Chalmers' arguments has left several conspicuous puzzles. Concerning HEC, since it is not clear that the cyberpunk implant realises cognitive processes, no instances of extended cognition have been established. Concerning HEM, Clark & Chalmers exhibit an apparent lack of conviction in the consequences of their own argument, but introduce *ad hoc* conditions to preserve their one hypothetical case.

---

[8] Clark & Chalmers 16–17, 17–18.

[9] Clark & Chalmers 17

[10] Clark & Chalmers 17. The parenthetical numbering is mine, following Sprevak's notation (ms 13).

Mark Sprevak's ambitious reply[11] to Clark & Chalmers might be understood to fill these lacunae. He bears both good news and bad news. His good news for Clark & Chalmers is that HEC is inescapably entailed by functionalism. Sprevak argues that functionalist theories can be organised according to how finely-grained their functional roles are drawn. Relatively coarse-grained functionalisms make no principled distinctions between, say, mental rotation and computer or cyberpunk rotation. Fine-grained functionalisms distinguish between neural and computerised implementations, but are objectionable because if they disallow non-neural implementations, they beg the question against conceivable forms of Martian cognition.[12] The bad news, however, is that functionalism entails only radical HEC, not the version of HEC that Clark & Chalmers articulated. Sprevak claims that the parity principle entails not only that the cases found questionable by Clark & Chalmers are indeed cases of extended cognition, but that there are even more radical cases: e.g. that contents of volumes in a library are beliefs of any person in the library, or that being in possession of a graphing calculator gives one a knowledge of integral calculus.[13] These consequences, Sprevak argues, are ridiculous, and serve as a *reductio* not only of HEC but, since it entails HEC, of functionalism.

Sprevak's conclusions, however, are too hasty. In order to show clearly where he goes wrong I will rebuild the case for HEC from the ground up. In section III I will sketch an account of the sort of modest HEC that Sprevak claims is impossible. Then in section IV, I will address Sprevak's arguments directly, illuminating how his misinterpretations undermine the effectiveness of his criticism. Before I proceed to my sketch, however, I will be useful to review some features of functionalism, and to formulate the distinction between computational and rational individuation of functional roles.

II. Functionalism

In order to remain as non-partisan as possible about the details of mental apparatus, I shall introduce the notion of an *item*. I shall use this term to denote such general things as objects, properties, and events. A *mental item*, for example, may be a mental entity, state, process, &c. By maintaining ambiguity between such kinds of referent I hope to avoid presuppositions about what kind of

[11] forthcoming

[12] Sprevak's arguments here draw broadly but unfaithfully from Clark (2005, 2008).

[13] Sprevak, ms 16–18.

ontology should ground the apparatus of mentality.[14] Likewise, terms such as 'physical item' are intended to obviate the appearance of such ontological presuppositions for physics, and so on. I am also concerned in the current discussion not to presuppose too simplistic a relation between mental items and what we wish to call *minds*. Certainly the apparatus of mentality, consisting of a distinctive architecture of mental items and mechanisms, might happily be called the apparatus of the mind. But because it is contentious to identify this apparatus with the mind *per se*, I will avoid speaking of the 'mind' as an entity, using instead the term 'mentality,' referring to a topic or collection of phenomena.

Some Functionalist Anatomy

Functionalism holds that the criteria of individuation for mental items are given solely by elucidating their 'functional roles'—their relations to mental and other kinds of items in a 'functional economy.' Functional economies are theoretical structures in which mental items interact in specified ways. These economies are traditionally thought to consume sensations or perceptions as inputs, and to produce actions, behaviour and beliefs as outputs, or something like that.[15] Functionalism is opposed to views on which the identity criteria for mental items are determinable independently of each other, or through their intrinsic properties or the intrinsic properties of their realisers. There are many varieties of functionalism, however, and I will take a moment to articulate some distinctions that will be relevant to my discussion of modest HEC.

Functionalists often identify themselves as 'causal role' functionalists, meaning broadly that mental function-types are posited and individuated based on the causal difference they make in a mental economy.[16] A familiar toy

[14] I also mean to steer clear of presuppositions about whether events are themselves simply a subtype of properties, as such presuppositions have potentially substantial consequences.

[15] This characterization is almost certainly problematic if read too crudely or taken too seriously (cf. especially Clark 2008, Noë 2004). In particular, many things easily called actions and sensations may be better classified as internal elements of the mental economy, than as entry or exit transitions of that economy. However, this gloss gives the right general idea.

[16] The 'role' in 'causal role functionalism' might, of course, indicate a commitment to role rather than filler functionalism, but I am not certain that all self-described 'causal role functionalists' mean this. Rather, the 'role' descriptor seems sometimes to indicate merely that causal role functionalists individuate mental items by their causal roles, not that they *identify* mental items with roles rather than fillers. At any rate, causal role functionalism is only mined here for a toy example, and the distinction between role and filler functionalisms will be held more conscientiously in view henceforth.

example is that pain is generally caused by bodily damage, and in turn tends to cause affective arousal, a desire that the pain stop, damage-avoidance behaviours, &c. This is not a particularly plausible functional specification of pain, but let us say that it is the total functional role of *toy-pain*. So long as this role obtains in an individual, then that individual is in toy-pain. So Jones is in toy-pain just in case she has been physically hurt, she is riled about it, she tries to keep her body from being damaged further, and she wants the whole episode to end. However, few theorists take *all* causal relations in the machinery of mentality to be constitutive conditions for mental items. Say for the sake of an example that in addition to its (actual) functional role, pain always has the effect that heat is generated in the cerebral cortex. Consider the physically-impossible but conceivable scenario in which an item occurs that satisfies the causal profile for pain *except that* it does not cause cranial heating; few would be comfortable casually denying that the item is mental, or that it is pain. If we do not include the generation of cortical heat in our specification of the causal role, however, then only a subset of the causal concomitants of a functional state constitute its functional role, and we should strive to have principled criteria for selecting which generalisations to include in the role and which to exclude. The problem of articulating functional roles becomes even more difficult if we aim to account not only for human mental economies, but those of non-human creatures or conceivable alien creatures. Theorists are sometimes willing to abandon plausibly constitutive elements of causal roles, e.g. that pain produces a desire that the pain stop.[17] This is all to say that while 'causal role functionalism' is a popular flag to fly, it represents a vague enough platform to admit of diverse interpretations, and its popularity does not reflect wide agreement among philosophers of mind about the methodology of functionalist enquiry or the substance of functionalist theories. In this section I will be concerned to distinguish two kinds of functionalist theories that employ different kinds of criteria for individuating functional roles. Both of these kinds of theories are orthogonal to causal role functionalism as such, but since causal role functionalism is so useful for examples, I will use the toy-pain example to illustrate one more simple but significant distinction between functionalisms.

Functionalists all hold that the work of individuating mental items is done by the notion of a functional role, but may identify mental items either with the role itself or with its occupant—its 'filler.' I will also refer to fillers of

[17] Cf. e.g. Lewis 1980. Though Lewis does not call himself a functionalist, he is a functionalist in my sense, which follows what Block (forthcoming) calls metaphysical functionalism.

functional roles as 'realisers.'[18] These two varieties of theory are usually called *role functionalism* and *filler functionalism* respectively.[19] Since filler functionalists identify mental items with physical items, they are sometimes called psychophysical identity theorists. However, an account is 'functionalist' in the sense I mean so long as the work of individuating items is done by a functional role.[20] To return to our toy-pain example, say (indulging another venerable philosophical simplification) that toy-pain is realised by the stimulation of the kind of nerves called 'C fibres'; that is, bodily damage causes stimulation of C fibres, and the excitement of C fibres in turn causes the kinds of arousal, desire and behaviour that are characteristic of toy-pain. Role and filler functionalists can agree that Jones is in toy-pain just in case her C fibres are excited; every time Jones is in toy-pain, one can point to either firing C fibres or the role in Jones' functional economy that they serve. Role and filler functionalists disagree about which one of those parts of the scenario is Jones' pain. Role functionalists about toy-pain identify Jones' toy-pain with the functional role that is satisfied for her when she hurts herself and gets upset and so on. If toy-pain items are properties, then Jones' toy-pain is the second-order property of having the property that is caused by bodily damage and of causing arousal and the rest, and she is in toy-pain if a part of her, such as her C fibres, fills that role. Filler functionalists about toy-pain, on the other hand, identify Jones' toy-pain with the activation of her C fibres. Again, if toy-pain items are properties, then Jones' toy-pain is the first-order property of her C fibres being stimulated. Theorists might be tempted to adopt either role or filler accounts of mental items because of concerns about the causal efficacy of mental items, and so on. Since these concerns are not central to my discussion, however, I will sidestep them as completely as possible.

---

[18] I choose this somewhat awkward terminology in part to conform to patterns in the literature, and particularly to Rupert's idiolect. I hope, however, to sidestep some serious concerns about the nature of realisation (c.f. e.g. Wilson 2001), the relation between realising and being a realiser (Rupert 2007), and so on. Though these issues are broadly relevant to my use of functionalism and to HEC, my discussion should be consistent with many ways of resolving such worries about the metaphysics of realisation.

[19] The popularisation of these terms is usually credited to McLaughlin, 2006. McLaughlin distinguishes token and type variants of each, but this dimension of variation won't concern me in this paper. The role/filler functionalism distinction may also be what Ned Block (forthcoming) is driving at in his distinction between metaphysical functionalisms that are conjoined with ontological functionalism on the one hand, and ontological physicalism on the other.

[20] What I am calling functionalism *simpliciter* is once more what Block (forthcoming) calls metaphysical functionalism, which is consistent with ontological functionalism and ontological physicalism (the conjunction of metaphysical functionalism with the former is role functionalism, with the latter is filler functionalism).

In order to remain as amenable as possible to both role and filler functionalists, I will strive to avoid making claims about mental items *per se* and instead focus on functional roles and their realisers.

Computational Functionalism

The most fruitful naturalistic programmes for developing more methodologically rigorous functionalisms have probably been the diverse work that describes mental apparatus as computational systems. I will refer to these views as species of *computational functionalism* or *computationalism* about mentality. Computationalist views are those on which the normal 'internal' relations of a mental economy can be modelled in terms of the algorithmic manipulation of representations. By adding the caveat 'normal' I mean to allow that some computationalist models may not claim to account for pathological or exceptional mental phenomena. By 'internal' I mean to refer to items that are internal *to a functional economy*, but not necessarily internal to the body or nervous system of an organism. Computationalism does not necessarily claim that items external to a mental economy or on its border (such as sensations or actions) can be modelled computationally. Finally, by using the vexed word 'representations,' I merely mean information-bearing items in something like Dretske's[21] sense. I do not mean to refer items that cannot be realised in a connectionist architecture, and I do not mean to refer to necessarily 'semantic' items that are compositional and error-sensitive and so forth. It should also be noted that what I call 'computationalism' is not what is called the *computational theory of mind*, which is usually limited in application to intentional states, and holds that intentional representations that are computed over are not merely information-bearing but semantic. Computationalism as I mean it applies equally to subconscious and introspectable mental apparatus, and does not attribute semantic properties to all representations. The contents of many computational representations will be what have sometimes been called 'non-conceptual' contents.

Computationalism informs a great deal of recent philosophical, psychological, and neurobiological enquiry, and almost all of cognitive science. It has hardly gone without comment, though, that the picture of our mental apparatus that is emerging from these various lines of investigation is deeply dissimilar to our pretheoretic picture of mental life. The items and apparatus that are posited by computationalists—visual edge-detection, subconscious

[21] Cf. e.g. Dretske 1986

sensorimotor models, spreading activation, body schemata and so forth—are not easily identified with or related to paradigmatically 'mental' items such as beliefs, desires, intentions, &c. Consequently, the word 'cognitive' has been appropriated to replace 'mental' for describing the objects and products of computationalist research programmes. Henceforth, then, I shall observe this custom and speak of *cognitive items* and their roles in a *cognitive economy*, &c. when referring to the posits of computationalist accounts. Computationalism is consistent with both *machine-state functionalism* and most forms of *psychological* or *psycho-functionalism*,[22] since computational models may or may not be based on a cognitive ontology of total state-descriptions.

Computationalism does not entail a particular account of the nature of cognition; it is merely a methodologically-significant articulation of what is widely held among cognitive theorists to be a necessary condition on cognition. Computationalists are certainly not committed to the view that the possibility of computational modelling is a sufficient condition on cognitive apparatus. Such a view would no doubt be radically permissive, entailing that cell phones are cognitive systems (and, for fans of a Wolframesque metaphysics of computation, that every physical system is a cognitive system). Since computationalism is merely an element of an account of cognition, it is consistent with many kinds of cognitive theories. Computationalists may hold that in addition to the possibility of computational modelling, cognitive items necessarily bear 'non-derived' content, [23] or subserve the flexible, adaptive behaviour of an organism,[24] or what have you.

Rational Functionalism

However, computational individuation is not the only methodologically rigorous strategy for specifying principled functional roles. Many functionalists individuate mental items based on rational criteria. Views broadly conforming to this approach might be called *rational functionalism* or *rationalism* about

[22] Specifically, almost all psychological theories that fall under the purview of cognitive science are computationalist. The models posited by, e.g., social psychologists, on the other hand, may be better described as rationalist models.

[23] Cf. e.g. Adams & Aizawa 2001, Searle 1980

[24] This is meant to express only the spirit of certain views including, perhaps, Clark's. This condition is probably not a sufficient adjunct to computationalism, for adaptive qualities of, say, trees may be modelled computationally though we would be biting a conspicuous bullet if we granted that trees think, or have cognitive processes. On the other hand, once we observe the distinction between mentality and cognition, even this claim may only *appear* to be an unpalatable concession.

mentality. Mental rationalists individuate mental items according to their roles in a rational economy, in which items with semantic contents stand in rational relations to each other. By 'semantic contents'[25] I mean that the contents of rational items are not merely information-bearing like the representations of computationalism, but have compositional structures, correspond to states of affairs of possible worlds (they are intentional), and so on. Rational relations are something like relations in which the content and kind of an item bear on the contents of other items. For example, *beliefs* are favourite toy items in rationalist ontologies. Judgments and beliefs may justify, or make unjustified, other beliefs in the same economy; inconsistent beliefs create pressure to resolve the inconsistency by abandoning or altering one or several of the implicated beliefs, &c. Of course, there is broad disagreement about both the nature of semantic items and of rational relations. There is also disagreement about how theoretical and practical rationality are related, how rationality and justification are related, and about whether either one is at bottom deontic, or instrumental, or mechanistic,[26] or something else. I do not mean to signal allegiance here to any particular view; any of these notions of rationality might subserve a rational functionalist account of mental phenomena.

Rationalist models of mentality are not common in cognitive science, but are standard fare in epistemology, moral philosophy, decision theory and philosophy of action. Rational functionalisms are typically tools for theorists who are frequently concerned with accounting for justification or prudential explanation.[27] Rationalist models of mentality are also common in the social sciences, e.g. economics, political science, sociology, &c., where investigators might not be concerned about justification *per se* but are interested in the interactions of agents that are sensitive to prudential considerations.

It behoves me to be clear about some things that rationalism is not. First, rationalism is not phenomenology, and since they have different criteria of individuation rational items are not necessarily items of which we are aware, or even possibly aware. We may never have introspective access to many or all of

[25] I could also have used the term 'propositional content' here; I avoid the term primarily because propositions are sometimes regarded merely as any structured content, and I wish to emphasise the difference between what I call semantic contents which interact rationally, and merely informational contents whose computational interactions are purely syntactic.

[26] I am referring here to certain theories of practical rationality that draw inspiration from Hobbes or Hume. Despite the fact that these accounts are sometimes styled as 'mechanistic,' they are not syntactically driven in the way that computationalist accounts are, for the operations of mechanistic accounts are still driven by the semantic contents of posited items.

[27] Though I certainly do not mean to suggest that e.g. cognitive scientists are unconcerned with notions of justification or fitness.

the items in our own rational economy, for the items of rationalist models are not necessarily accessible to consciousness or distinguishable according to qualitative feelings. Second, many accounts called 'analytic functionalism' may be kinds of rational role functionalism, but I prefer to distinguish models not according to the kinds of enquiry and evidence that they permit themselves, but specifically on the individuational criteria they permit themselves. Insofar as my method of categorising functionalisms cross-cuts other taxonomies, I am happy to abandon correspondence with them.

Finally, rationalist accounts of mentality are not folk psychology, where by folk psychology I mean to refer to the naïve theories of mind generally employed by people (but not people qua theorists) in the course of social interaction. There are often similarities of ontology and mechanism between rationalist accounts of mentality and folk psychology, probably because much of the mechanism of folk psychology is based on rational interactions of mental items. However, folk psychologies usually involve the application of mixed models, on which some of the mechanisms of mentality are rational, but many are merely causal. Furthermore, since rationalist models are used as theoretical tools they are often more principled, sophisticated, and free from contradiction than folk psychology. It is also the case that many rationalist models bear only strained resemblance to folk psychological views; for example, Robert Brandom's semantic inferentialism is a clear case of a rationalist theory, but trades in items like 'doxastic commitments' and 'material substitution-inferential commitments' which have no place in folk psychology. Other rationalists have posited mental items called pro-attitudes,[28] acceptances,[29] understandings, [30] &c. that are not natural elements of folk theories. Nevertheless, rationalist models of mentality bear a much more comprehensible relation to folk psychology than computationalist ones. While not all rationalisms employ familiar folk psychological items like 'beliefs' and 'desires,' many do make use of such notions, or very similar ones. Indeed, successful rationalist models might be viewed (and have been by some authors) as highly-refined folk psychologies.[31] Accordingly, I will take rational models of mentality to express our best accounts of commonsense psychological terms such as 'belief' and 'intention.'

---

[28] Cf. e.g. Davidson 1963

[29] Cf. Cohen 1989

[30] Cf. Schick 1991

[31] I am thinking of e.g. Sellars (1962). Indeed, insofar as theories employed as folk psychologies are learnable or culturally mutable, rationalist models could in principle come in part or in whole to supplant contemporary folk psychological theories.

Computational Roles and Rational Roles

I have sketched two categories of functionalism called 'computationalism' and 'rationalism' that employ distinct vocabularies for individuating items, and whose associated research programmes seem to have produced distinct bestiaries of items and mechanisms. Nevertheless, despite their differences, it is common to suppose that computationalist and rationalist models describe the very same apparatus, only at different levels of detail. It is clear enough, after all, that human cognition and rational mentality are related in subtle and intimate ways. I wish to suggest, however, that despite their complex interrelations computational roles and rational roles are plausibly distinct kinds of posit, varying in more than just detail. The upshot of my claim is that arguments and generalisations pertaining to one kind of functional model may not pertain to the other.

To begin with, it is not controversial that computation and rationality are distinct notions, with distinctive instances of application. Consider, for example, a pocket calculator, which is certainly capable of computation (though not, in normal circumstances, computation that would qualify as cognition). The functional economy of a calculator can be described in terms of the computational roles of its electrical states and state-transitions, and the total architecture of these roles can be described as an algorithm. Yet a calculator is not capable of rationality. It and its internal states are inappropriate objects of rational judgments: it cannot be irrational; that it displays '42' on a particular occasion cannot turn out have been a bad idea for the calculator. It is not intelligible to ask of an item in a calculator's functional economy whether it is justified, or prudent given the calculator's aims. One might ask such questions about elements of the calculator's design, but not about the course of its operation. Similarly, consider a computational model (as sophisticated as you like) of visual object recognition as implemented in humans from the retina through the ventral stream. The neural mechanism described by such a model that enables (say) Jones' visual recognition of her coffee mug can be categorised in terms of the computational roles played by neurological items and so on, but it is still daft to ask whether the whole process is rational or prudent. Even if some physical defect in Jones' brain led to a malfunction, causing deviation from the normal cup-recognition algorithm, the defect of computation is not a defect of rationality. One might ask whether human object-recognition is computationally efficient or how well it promotes biological fitness, but such questions go beyond the operation of computational models and still are not obviously questions about rationality. On the other hand, it is perfectly

intelligible to ask whether, given her circumstances, it is rational for Jones to believe that her mug is in front of her. All this is to say that computational criteria as such are genuinely distinct from rational criteria.

The calculator and visual-processing examples illustrate that mental items that serve in the operation of a computational economy and those that have roles in a rational economy are subject to different predicative attributions. That is, rationally-individuated items have particular rational properties (e.g. 'is justified' or 'is imprudent') or have the property of intelligibly being subject to rational scrutiny. Leibniz' Law suggests that since computational and rational items have divergent properties, they must not be the same items, but this is misleading. Of course it is true that the set of all computational representations and processes is distinct from the set of rational contents and relations. Computational models, after all, manipulate merely information-bearing (largely non-conceptual) representations by stipulation, whereas rational models must manipulate compositionally-structured items with intentional contents. Computational models specify algorithmic state-transitions whereas rational models avail themselves of rational relations. Still, all these stipulated differences are consistent with the possibility that the rational items are identical to a principled subset of computational items that are not *merely* information-bearing, and to which the special properties attributed generally to rational items are applicable. Computational models might just be more detailed than rational models.

Moreover, the view that rational items just are a subset of computational items is an alluring one. Part of the promise of computationalism is its application to the mind-body problem. We understand how computers can be physically implemented in silicon-based electronics and other sorts of physical systems, after all, and if some particular intricate computational architecture can account for rational relations, then whatever it is that rationalist functional models describe might be understood as an un-mysterious physical process. I do not deny the significance of this research programme, nor am I claiming that mentality is, after all, something spooky. Still, the mind-body problem cannot be assumed away. Even if rational mentality can be modelled computationally, as I expect it can, this possibility is not sufficient to establish that rational roles are identical to a subset of computational roles.

Imagine, for example, that there is a complete computational theory *C* of Jones' cognitive economy that predicts, say, her actual behaviour within some standard of accuracy given a complete enough description of her physical environment. Imagine also that there is a similarly complete rationalist theory *R* of Jones' beliefs, desires and so forth that accurately predicts her actions given

her environment. Now, suppose that Jones wants to drink coffee, and this is represented in *R* by a rational functional role *Ψ*. That role is specified in terms of rational relations; perhaps Jones believes that her coffee mug is in front of her and is full, in which case (other things being equal, assuming no contravening considerations and so on) she will take the action of picking up the cup and drinking from it or whatever you like. Jones' desire is also represented in *C* by a functional role *Φ*, which is specified as an intricate algorithmic structure with variables set thus and so. Both *Ψ* and *Φ* are realised in Jones' brain by some pattern of neural activity. If the elements of *R* correspond in principled ways elements of *C*, then the realiser of *Ψ* and that of *Φ* may be numerically identical; both realisers are just the aforementioned pattern of neural activity. So a theorist committed to computational filler functionalism and to rational filler functionalism would only be committed to the desire being a single cognitive/mental item in Jones.

Nevertheless, the *roles Ψ* and *Φ* may still be distinct in this case. To begin, they are articulated in distinct vocabularies, which confer on their roles distinctive logics of functional interaction. *Ψ*, as a rational role, may justify or obligate and so on. *Φ*, as a computational role, has formal relations in an algorithmic architecture. It would not be trivial to demonstrate that these vocabularies are fully inter-translatable.[32] One might suspect that if they share a realiser, some argument from the transitivity of identity can establish the identity of the roles, but no such simple argument exists. Roles cannot be identical to their own realisers since second-order items are not identical to the first-order items in terms of which they are specified. That is, if *Ψ* (a role) were the property of having the property (a filler) that is caused by *X* and causes *Y*, then *Ψ* cannot be identical to the property that is caused by *X* and causes *Y*.

So how might *Ψ* and *Φ* be related? It may be that they merely correspond, meaning that they share a physical realiser, but are otherwise unrelated. [33] Alternatively, *Ψ* may stand in the relation to *Φ* that *Φ* stands in to its physical realiser. That is, rational roles might be realised by computational roles. Rational roles might even commonly be multiply realisable with respect to computational roles. A third option is that some or all rational roles might turn out to be

[32] I am not invoking anything like conceptual schemes here (cf. Davidson 1974). Functional models are *formal* objects, so inter-translatability is no trivial thing even for the committed Davidsonian.

[33] This situation might be in the spirit of Donald Davidson's anomalous monism (cf. 1970), but with computational roles substituted for physical events. That is, this would be anomalous corealisation, with the differences between the individuation criteria for computational and rational roles somehow assuring the impossibility of type-type correlations.

identical to computational roles. I do not think that the answer to the question of how rational and computational roles relate, if there is a unique answer, has been settled. After all, most metaphysical issues related to the mind-body problem remain controversial.

I am not claiming that rationally-individuated and computationally-individuated functional roles are actually distinct. However, so long as it is unclear how they relate to each other, certain metaphysical assumptions will turn out to be unmotivated. In particular, it cannot be assumed that claims about computationalist theories generalise to rationalist ones, or vice versa.

III. Modest Cognitive Extension

In this section, I will sketch a defence of cognitive extension based on Clark & Chalmers' argument. My aims are primarily clarificatory, so I am only concerned to defend a weak (I prefer 'modest') version of HEC, as free as possible from unnecessary frills and commitments.

Concerns about Parity

Since the parity principle was the backbone of Clark & Chalmers' argument, we should take care to be precise about what it means. First of all, the parity principle is a normative rule for how we *should* organise our cognitive theories, not a generalisation about how in fact we do.[34] Second, the cognitive theories in question are functionalist theories. Since the parity principle refers to cognition rather than mentality, we are looking at computational functionalisms.[35] Third, the parity principle should be distinguished from the claims that Clark & Chalmers use it to support, namely HEC and HEM. Fourth, I take the parity principle, and consequently HEC and HEM, to be concerned with the locations of the *realisers* of cognitive items rather than the locations of cognitive roles or cognitive items themselves. Clark & Chalmers write that the parity principle is about locating cognitive processes which seem to be, in my terminology,

[34] Clark & Chalmers are quite clear throughout their 1998 that this is their game, and Clark reiterates this in his 2008 restatement (p. 77).

[35] Clark & Chalmers do not explicitly link their original argument to functionalism, though they are suggestive: they invoke the notion of a function when discussing the parity principle and HEM (Clark & Chalmers 8, 13). Nevertheless, functionalism is definitely in the water and has been invoked by Clark in defence of Clark & Chalmers' original claims. Cf. e.g. Clark 2008, pp. 88–89, though I will conspicuously part company with Clark on the strategy he articulates in that passage, since it only obviously concerns HEM, and not HEC.

cognitive items. However, more recently[36] Clark has insisted that arguments about parity are meant to apply to the vehicles of cognition, rather than the contents.[37] Such equivocation between cognitive items and realisers is fine for filler functionalists, but is vexing and possibly objectionable for role functionalists. Following Clark's more recent hints, I will take Clark & Chalmers' claims to be about realisers and not roles or items as such; I will not concern myself with justifying or defending claims that cognitive roles or cognitive items are extended. Such a revision of the scope of Clark & Chalmers' arguments is in contrast to interpretations such as Rupert's, who takes HEC to involve the claim that body-external items are proper parts of cognitive states.[38]

Even clarified thus, however, there are some open questions of interpretation. Robert Rupert imagines that the parity principle could mean one of two things: either that whether a realiser is cognitive does not depend on its absolute location, or that whether a realiser is cognitive does not depend on its location at all, even relative to other items. The second interpretation is based on a natural reading of the original passage by Clark & Chalmers that seems to suggest that we should call an item a cognitive vehicle, so long as we would call it a cognitive vehicle if it were in a head. Rupert dismisses this claim as assuredly false. It is implausible, he suggests, that an item that is normally a paradigmatic cognitive realiser, once shorn of its normal causal relations to other items, remains cognitive. He considers the case of a single neuron, which would almost certainly play some role in a cognitive system were it in the brain of a healthy animal, but which plays no such role outside of that context. For example, if the neuron were isolated from other neurons and in a saline preparation in a petri dish, it is doubtful that it would play any recognisably cognitive role.[39]

Since the second interpretation is false, the parity principle's best chance is on the first reading, on which it just entails that 'cognition is cognition, *wherever* it occurs.'[40] This, Rupert claims, is certainly true but of questionable dialectic value. If we had a theory of what cognition is, we presumably would not need the parity principle to help us find instances of it. Rupert's final line

---

[36] Cf. Clark 2008, p. 77

[37] I confess that I am not certain I know what the vehicle/content distinction refers to. Susan Hurley's well-known remarks seem to suggest that it refers to the distinction between an apparatus and its states, or an algorithm and its values. Clark's gloss (2008, p. 76), on the other hand, is suggestive of the distinction between a realiser and the item it realizes. Since I am more concerned in this paper with Clark than Hurley, I will interpret the term 'vehicle' in this latter sense.

[38] Cf. Rupert 2004, 389

[39] Rupert forthcoming-a, 5–6

[40] Rupert forthcoming-a, 5

on the parity principle is that his 'reasoning does not demonstrate the falsity of [the parity principle], so much as it shows that we can have little confidence in its naïve application.'[41] This is precisely the worry that undermined the discussion of Clark & Chalmers' Tetris thought experiment in section I. An account of the necessary and sufficient conditions for cognition might always undermine the claim that a putatively extended realiser[42] of a cognitive role can really be cognitive. If Rupert is right, then the parity principle is at best useless.

Rescuing the parity principle from objections like these is not easy. Clark's views are articulated in fragments throughout his oeuvre and probably have transformed somewhat since he and Chalmers first wrote about the parity principle more than ten years ago. Therefore, I will replace the principle with a novel treatment, consisting of two theses and a material rule of inference. These will be subject to the four clarifications of the parity principle articulated at the beginning of this section: like the original parity principle, they are prescriptive, assume computational functionalism, are distinct from HEC and HEM, and concern cognitive realisers. Although my discussion is closely based on Clark & Chalmers, I will not be overly concerned with remaining faithful to their views.[43]

The Parity Principle Restated

The first thesis follows from functionalism, and might be called the *location indifference thesis*. The indifference thesis just states that the skull (or the boundary of the central nervous system or whatever) does not mark a principled region inside of which all cognitive vehicles must necessarily (in the sense of logical necessity) be located. In fact, no physically-identified boundary necessarily coincides with such a region, since (according to functionalism) functional realisers are not identified according to physical criteria, but according to functional criteria.

---

[41] Rupert forthcoming-a, ms 7

[42] Since the realisers of extended cognitive items are not entirely outside the brain or body, it would be infelicitous to use the term 'external realiser' to refer to them. I will use the phrase 'extended realiser' to refer to these items.

[43] Even so, although I will depart from the letter of Clark & Chalmers' arguments and despite Clark's endorsement of the parity principle as such, I do not believe that my interpretation is a departure from the spirit of Clark & Chalmers' view.

> Location Indifference:
> There are no logically necessary physical restrictions on the locations of realisers of cognitive roles.

This thesis is identical to Rupert's first interpretation of the parity principle. Location indifference is accepted by Adams & Aizawa[44] and Rupert[45], all of whom take it to be entailed by functionalism and none of whom accept HEC. Their arguments reveal that this thesis alone is not sufficient to establish HEC, but I will take it as the first step in a line of argument that can support HEC.

The second thesis that I associate with the parity principle expresses a methodological claim about how to treat evidence, and might be called the *cognitive conservation thesis*. 'Conservation' is the claim that because a function-type can be realised in a non-biological or, indeed, non-cognitive medium does not mean that it is not sometimes a cognitive function:

> Cognitive Conservation:
> If a functional role φ is realised by an item p and p is not a cognitive realiser, this fact is not evidence that function φ is never a cognitive role.

Another way to express the conservation thesis is in terms of *functional parity*, a term of art that I introduce here:

> Functional Parity:
> There is said to be *functional parity* about a functional role φ between an item p and an item q if and only if p realises functional role φ and q realises the same functional role φ.

Functional parity, or simply 'parity,' is not a remarkable relation. Functional realisers may have parity with other realisers of roles in the same functional system, or in other systems of the same system-type, or even across system-types. And parity may obtains in non-cognitive functional systems (e.g. internal combustion engines, televisions), as well as cognitive systems. Since parity is cheap it cannot buy much, and certainly not a claim as theoretically dear as HEC. But it does not buy cognitive contraction, either. What cognitive conservation entails is that if parity obtains between a head-internal and a head-external item, that fact by itself should not lead us to change our judgments about whether the

[44] E.g. in their 2001, though the version they endorse is what Rupert calls a 'weak modal claim,' as discussed in the text below.

[45] Rupert (forthcoming-a) wanly endorses his 'vacuous' reading of the principle, but does not take it to support HEC.

functional role can be a cognitive role. In effect, cognitive conservation holds that functional parity with head-external items should not be evidence that a head-internal item is not a cognitive realiser.

Functionalists should accept conservation, since it is just a special case of the view that whether a functional role is cognitive or not should depend on the place of the functional role within a particular kind of functional economy, and not on the kinds of items that possibly realise it. Suppose that there is some computational specification Φ of an edge-detection process that describes an element of human visual processing and an element of a particular high-tech fingerprint-analysis program. Conservation entails that parity between these implementations of Φ would not be evidence that a part of human visual system is not cognitive, after all.

The need for a thesis like 'conservation' is motivated by functionalism and indifference. If we are given an item such as a cluster of neurons that is likely to realise a cognitive function, and an item such as notebook that is not likely to realise a cognitive function, and we are asked at some point to treat them as vehicles of the same functional role in a cognitive economy, should it be a cognitive functional role or not? Assuming that the right kind of parity obtains,[46] indifference demands that we not call one token of a function 'cognitive' and not the other based only on the physical characteristics of the realisers. Yet one might accept indifference, and yet use parity to argue for the *contraction* rather than the growth of the realisation base of cognition—arguing that if a head-internal item functions like head-external items, it does not really realise a cognitive item after all. Herbert Simon[47] proposed just such a cognitive contraction for biological memory, reclassifying it as external to cognition, though realised inside the brain. There are several practical reasons not to take Simon's route here.[48] Even if one were compelled by a view like Simon's, however, it should not be the fact of parity that justified such a view, but because the functional roles associated with memory fail to satisfy requirements on cognitive functions for some independent reason. If contraction in the style of Simon is not an option, then, then the realisation base of cognition must *expand*

[46] What exactly I mean by 'right kind of parity' will be specified below, in my discussion of the parity rule.

[47] Simon 1982

[48] Clark (2001, p. 158 in response to Simon 1982, p. 65) accuses of Simon of misunderstanding the extend to which biological memory is actively and dynamically organised, and of relying on a conception of cognition too tied to a flawed concept of agency. Haugeland also dismisses Simon's theoretical strategy (1998, pp. 210–211).

whenever indifference demands that we treat the neuron and the notebook the same.

The case of edge-detectors that implement Φ is not the right kind of parity for cognitive extension, though. Although the neural and software realisers of Φ *ex hypothesi* implement the same algorithm and therefore have similar functional specifications, there are two reasons to treat them differently. First, although these instances of Φ share internal functional structures, they may have divergent roles within their respective economies. Outside of Φ, the gross functional architecture of visual cognition and fingerprint analysis are no doubt quite different. If some Martian visual system included a fingerprint-analysis module that shared the structure not only of Φ but of the entire fingerprint-analysing software programme, we might imagine that this module would cease to be considered cognitive if its realisers were isolated from the rest of a cognitive system (e.g. the neurons removed from a brain and placed in preparation). Surely, then, the functional superstructure containing a role, more than the inner structure of the role, should be our guide when we attempt to determine whether the role is cognitive.[49] Even a neurological structure such as the mirror neuron system, which is activated for different purposes in different contexts, does not always fill the same functional role each time it is activated within the same token cognitive economy.

Second, instances of Φ may be said to have divergent properties because the economies of which they are parts have different properties. Recall that computationalism is based on a necessary condition that cognitive systems have computational structure, but that this is not a sufficient condition. The software programme probably does not satisfy other requirements for being a cognitive system—it does not have an appropriate role in guiding the behaviour and bodily regulation of an organism, or operate over non-derived representations or what have you.[50] Any contentious case for cognitive extension must observe these two considerations. It must show not only that putative brain-external realisers of cognition have *some* kind of parity with brain-internal realisers, but

[49] This is a jab at criticisms by Adams & Aizawa (2001) and Rupert (2004), that parity does not obtain between, say, (T1) and (T2) or between Inga and Otto, because the fine structure of functional implementation diverges and therefore the functional roles of internal and putatively extended cases are distinct. Since Clark (2008) offers a satisfactory response to these objections, however, I will not be concerned in this paper to deflect them. This is also the criticism that Sprevak (forthcoming) is concerned to respond to in the first part of his discussion.

[50] This is part of the reason that the realisers of a simulation and the items they simulate need not have all the same properties, although they are functionally isomorphic and therefore have parity.

that they occupy the same position in the gross architecture of the entire functional system, where that functional system is a cognitive system.

And there is another caveat. In making the case for cognitive extension we must also distinguish, within the context of a cognitive system, which functional roles are cognitive and which are not. Playing a functional role in a cognitive economy should not be sufficient for being a cognitive functional role. After all, items such as the structure of ambient light and the physical properties of muscle tissue play functional roles in the context of human cognition. Ambient light focused on the retina enables the first steps of visual processing, and the properties of muscles contribute to the translation of neural action potentials into bodily movement. These are important aspects of the input- and output-functions for cognition. But cognitive theorists should balk at the notion that such functions are cognitive merely because they figure in the theory. A bubble rising in boiling water is not an instance of a cognitive process just because someone saw it.

HEC-theorists need not claim, however, that all such items on the fringe of a cognitive economy are cognitive items, or even candidate cognitive items. Clark & Chalmers explain that in the cases in which the parity principle is supposed to apply, 'the relevant parts of the world are *in the loop*':[51]

> In these cases, the human organism is linked with an external entity in a two-way interaction, creating a *coupled system* that can be seen as a cognitive system in its own right. All the components in the system play an active causal role, and they jointly govern behaviour in the same sort of way that cognition usually does.[52]

In other words, Clark & Chalmers claim that some items that are realised outside the brain or even outside the body may be more happily thought of as realising *intermediate* items in a cognitive economy than as realising border items. It is *these* extra-cranial or extra-bodily items that are strong candidates for realising extended cognitive functions.

Nor is every function realised in the head is a cognitive function. A random mereological sum of head-internal items—say, a cluster of dendrites and some synaptic potentials scattered across the left temporal lobe—is not a candidate for realising a cognitive functional role if it does not realise a particular functional role according to our theory. If an item does not realise a cognitive functional role it cannot support inferences to cognitive extension. Say an inert

[51] Clark & Chalmers 9

[52] Clark & Chalmers 8

tumour or a rock were lodged in Jones' brain (poor Jones!). If the rock's presence makes no difference to the normal goings-on around it, it plays no functional role and therefore cannot be the basis of extensions of the realisation base of cognition. Jones' rock then cannot justify the inference that rocks are cognitive realisers in general, or even that rocks are parts of Jones' cognitive economy, since it has no functional role about which any parity can be identified. If the stray rock in Jones' head did come to make a difference to Jones' cognitive profile, say because Jones' neural networks became accustomed to its presence and used its properties to realise some computations, other rocks (even very similar rocks) would still not be candidates for realising extension. Other rocks could have the right kind of parity with Jones' stray rock only if they somehow came to play the very same role that Jones' stray rock plays in her cognitive economy.

Clark & Chalmers' proposed test for identifying strong candidates for cognitive extension is what I call the *parity rule*, which is a material rule of inference that goes something like this:

> Parity Rule:
>
> P1. item p realises functional role φ in a cognitive economy e of an organism s.
>
> P2. if some item q that realised functional role φ in e were located in the head of s, φ would be an uncontroversially cognitive functional role.
>
> C. item p is a cognitive realiser.

Functional roles, in this context, must not merely be articulable; the φ-role must be one that does explanatory or causal work in a true functionalist model.

While the second premise (P2) of this parity rule echoes the structure of the parity principle, it should be understood that three conditions on the functional role φ are built into it:

> (P2.i) φ must be a functional role relative to a cognitive system e of an organism s.
>
> (P2.ii) φ must be a *cognitive* role relative to cognitive system e.
>
> (P2.iii) φ must sometimes be realised by items inside the brain of s.

(P2.i) and (P2.ii) each prevent the parity rule from running afoul of one of the two caveats mentioned above. (P2.iii) is a more dialectic consideration, ensuring that the parity rule begs as few questions as possible. The extended realisers identified through the application of the parity rule need not be all of the extended realisers; depending on one's favoured brand of functionalism, one may

be willing to admit more or fewer cases of extended cognition in addition to the cases that satisfy the parity rule. Clark & Chalmers, for example, seem quite willing to at least entertain the possibility of rather rampant extensions.[53]

This version of the argument from parity is relatively conservative, and the parities that satisfy the premises—what I will call the 'right kind of parities'—are quite restricted. But it is appropriate that the initial case for HEC be made with the use of conservative assumptions. The parity rule itself does not suppose a particular account of when a functional economy is a cognitive system; it is merely a methodological 'rule of thumb'[54] that serves as a measure for the identification of extended cognitive realisers. Still, if one accepts functionalism—and thus, the location indifference thesis and the cognitive conservation thesis—one should accept the parity rule. The dialectic goal of this rule is just to get the foot in the door for cognitive extension.

Cognitive Impartiality

Before introducing the basic case for HEC, however, I should clarify what I take to be the strength of the claim. HEC is often understood in contrast to (merely) embodied cognition, which is roughly the claim that extra-neural bodily tissue can realise cognitive items. Rupert observes that HEC, as it is currently understood by its most conspicuous advocates and critics in the philosophy of cognitive science, should not be thought of as merely a weak modal claim that cognition is extended in some possible world. Rather, arguments about HEC are deployed in a dialectic space with a strongly empirical flavour, and which is concerned with revisionist attitudes in cognitive science. Therefore HEC should not be thought to be the 'weak modal' claim that cognitive extension is merely conceivable, or even that human cognition could be extended in exceptional circumstances or the cyberpunk future. Rather, it should be understood as the claim that human cognition is in fact frequently extended. [55]

> The Hypothesis of Extended Cognition
>
> Human cognitive functional roles are actually and commonly realised in part by entities outside the body.

[53] Cf. e.g. section 5 of Clark & Chalmers, 16–18.
[54] Clark 2008, p. 77
[55] Rupert 2004, 392

Clark draws attention to research by Wayne Gray, Wai-Tat Fu and colleagues[56] on what he calls the phenomenon of *cognitive impartiality*.[57] According to the framework employed by Gray & Fu, *hard constraints* are limitations on an organism's possible modes of interaction with its environment, whereas *soft constraints* are factors that might promote the use of certain possible modes of interaction over others. For example, withdrawing cash from an ATM is a process that requires a person to maintain a sustained interaction with the ATM. When engaged in this process a person faces hard constraints that determine the specific sequence of button-presses, &c. that must be performed in order to successfully complete the task, and soft-constraints that encourage, say, a particular pattern of saccades, or the use of internal resources to remember information (say of a PIN) rather than external reminders (a slip of paper with the PIN written on it).[58] Gray and colleagues report experiments in which participants are given interactive tasks, such as programming a VCR or reproducing an arrangement of coloured tiles,[59] and where the soft constraints of these tasks are manipulated across conditions. Gray et al. report that when participants can use either of two interactive strategies for information-retrieval—accessing the contents of their internal short-term memory, or sampling information in their visual field—the only determinant of their choice of interactive strategy is an optimisation function that is responsive to the soft constraints of the scenario. In Gray et al.'s studies the optimisation function was to maximally reduce the time-cost of interactive strategies. The participants were indifferent as to whether their interactive strategy involved all brain-internal or some brain-external realisers; their choice was determined only by which strategy got the job done faster.[60] Clark speculates that in other circumstances the optimisation function may well be determined by something other than time[61]—for example, if there were no time-pressure to complete a task and the

[56] Cf. Gray & Fu 2004, Gray et al. 2006.

[57] Clark 2008, pp. 118–122.

[58] Gray & Fu 2004, 361

[59] Gray & Fu 2004 report the VCR experiment. Gray et al. 2006 report the latter experiment, which is based on Dana Ballard's 'blocks world' task (Ballard et al. 1995, Ballard et al. 1997).

[60] Gray & Fu 2004, Gray et al. 2006.

[61] Clark 2008, 121, though this claim is in contrast to Gray et al.'s 2006 articulation of the Soft Constraints Hypothesis. Clark is right that by limiting the variation in the contexts of interactive behavior they investigate, Gray et al. have overstated the case for time-optimisation of interactive strategies. This is why Clark and I present Gray et al.'s data as evidence for what Clark calls 'cognitive impartiality,' which takes the process of soft-assembly to be guided by a contextually-determined optimisation function, rather than what Gray et al. call the 'Soft Constraints Hypothesis' which takes soft-assembly to be guided by time-optimisation only.

cost of error were very high, perhaps participants would make consistently more frequent and redundant use of more accurate perception-based strategies over less accurate memory-based strategies.

Here is a way of thinking about the consequences of cognitive impartiality for functionalist theories of cognition. Many kinds of sustained interactions of organisms with their environments admit of multiple strategies of interaction. Some element of the cognitive system of the organism must select a particular strategy to deploy when the organism is engaged in such interactive tasks. Gray et al. call this element a 'control system.'[62] This control system should not be confused with the parts of a cognitive system that correspond to faculties of agency, conscious decision-making and so forth—the control system that figures in Gray et al.'s model is a subconscious functional item which, though it is possibly overridden by higher-level decision processes some of the time, normally operates unnoticed by the person it serves. The behaviour of this control system is guided by contextually-determined soft constraints. Gray & Fu present data about sustained interactions between people and a VCR programming interface, examining what happens when the control system is faced with the choice between strategies for retrieving information. On the *conventional strategy*, the information can be encoded into short-term memory and retrieved as needed through the computational processes that are involved in such episodes. On the *extended strategy* the information can be left unencoded in short-term memory, and retrieved as needed through saccades and visual processing. The conventional strategy employs functional roles that are all realised inside the brain, but the extended strategy does not. Some elements of the extended strategy, such as motor events like saccades and the sensory events preceding visual processing, are normally at the fringe of a cognitive economy. The role of maintaining an accessible source of information that persists during the performance of the task is realised wholly outside the body in this case, by the independent persistence of a display in the visual field.

Gray et al. are investigating what kinds of soft constraints guide the behaviour of the control system that must select one of these strategies. One constraint might conceivably have been whether the functional roles implicated in information retrieval are all realised in the brain or not. Gray et al. show that there are cases where this difference is not relevant to the behaviour of the control system, and if their framework for thinking about the control of interactive strategies is right, the location of functional realisers is *never* relevant. It would be functional differences between strategies—the time it takes to

[62] E.g. Gray et al. 2006, 462

deploy a strategy, the accuracy of the information retrieved, or what have you—and not facts about where roles are realised *per se* that guide the behaviour of the control system. From the point of view of the cognitive system, it is absolutely immaterial whether the realisation base of a functional role is entirely internal or extended. Sometimes a computational apparatus, such as the apparatus implicated in realising the conventional strategy, is assembled out of all neural components. Sometimes, however, a computational economy is guided by soft constraints to exploit the computational capacities of extra-neural components; Clark calls such ephemeral computational alliances *soft-assembled systems*.[63]

It is in just such cases, where a cognitive system itself is blind to whether realisations of roles are internal or external, that the parity rule applies. The two strategies do have different detailed functional structures: one involves the use of short term memory and the other involves the use of perceptuomotor abilities. Nevertheless, they occupy equivalent roles in Gray et al.'s functionalist model of interactive behaviour, and I propose that the evidence for this equivalence is their *intersubstitutability*: if, within a particular cognitive economy, two distinct functional processes are selectively called by operation of a functional control structure and are intersubstitutable with respect to the completion of some cognitive task, then there is reason to consider the two functional processes distinct kinds filling a *single functional role* relative to that task.[64] That their fine functional descriptions diverge just means that there are distinct strategies that fill the general role on different occasions; when either of the strategies is employed on a particular occasion, it can satisfy the general functional role of retrieving a particular piece of information (e.g. 'what time *Star Trek* comes on' in the VCR task) that will be used to guide the organism's performance in an interactive task. One of the conditions on this general functional role is that the control system decides how to fill it based on the soft constraints it is fed. If Gray et al.'s model is correct, then, the role is a single, well-defined role implicated in the performance of a certain class of cognitive tasks. Furthermore, it satisfies all the conditions discussed above in connection with P2 of the parity rule. The role (P2.i) is a functional role relative to a particular cognitive system (any normal human cognitive system). It (P2.iii) is sometimes realised entirely inside the brain, and in such cases it (P2.ii) is a cognitive role. Applying location indifference, we should treat the role as cognitive no matter where it is realised.

---

[63] Clark 2008 p. 116

[64] If this condition is met, it might also do the dialectic work that Rupert supposes must be done by what he calls 'generic kinds' (2004, 418–421). The existence of such roles, e.g. in the context of interactive tasks like VCR programming, is a counterexample to Rupert's argument that such kinds are not actual.

Sometimes the role is realised by functional structures normally associated with short-term memory, but sometimes it is realised by structures normally associated with perception. *In just such cases*, when perceptual structures are called on by such a control system in aid of interactive tasks, those perceptual structures have the right kind of parity with cognitive memory structures. We should therefore recognise those structures at those times as cognitive structures. When they are implicated in the service of a cognitive role they should be considered not as boundary processes of a cognitive economy, but as fully *internal* processes of that economy, though not realised entirely inside the brain.

The Tetris case discussed by Clark & Chalmers can be thought of as an instance of the same kind of cognitive extension. As Tetris players become increasingly experienced, a larger part of the computational burden of game-playing is shifted onto on newly-trained unconscious mechanisms. Experienced players, then, come to cede certain decision-types to subconscious control systems. Perhaps one of these decision-types is whether to employ either a conventional strategy of mental rotation or an extended strategy of computer-assisted physical rotation. Since increasing time pressure is the primary driver of difficulty in Tetris, the optimisation function for this control system will be (as in Gray et al.'s work) guided by a policy of minimising time-costs. The reliable use of the extended computer-assisted rotation strategy that Kirsh & Maglio observed in experienced Tetris players might be the result a well-tuned optimisation function. If mental rotation and computer-assisted rotation are alternative strategies in this sense, then they have parity about a theoretically-significant functional role that is cognitive for a particular system, and is sometimes realised in the head.

Rupert's Objections

Rupert is not sanguine about this line of argument, however, and articulates three objections to Clark's use of Gray et al.'s research. Rupert's main objection is that he does not see how Gray et al.'s research supports HEC at all. Rather than being evidence that a cognitive system can extend beyond the body, they might simply show that 'when there is not great cost in terms of time, the cognitive system uses resources beyond its boundary.'[65] This interpretation of the results, however, begs the question quite completely. The parity rule is what justifies the use of an extended strategy as evidence for the genuine extension of the cognitive system; if computations fill a functional role that is a cognitive role

[65] Forthcoming-a, ms 32

for a particular system, the parity rule demands that they be considered cognitive computations even if they are sometimes realised outside the brain. What Rupert is suggesting here is just that we might think of the computational contributions of body-external objects as merely causal and not cognitive, although the same computational contributions are considered cognitive contributions when realised in the brain. Rupert charges that the serendipitous forays into the environment that Clark's 'cognitive systems' make are unprincipled, but an appreciation of the parity rule returns the ball firmly to Rupert's court.

Second, Rupert objects to Clark's endorsement of the claim that 'the central controller makes no functional distinction between knowledge in-the-head versus in-the-world.'[66] Rupert claims that 'Gray et al.'s full model *must* draw a functional distinction' since one strategy uses perceptual apparatus and the other does not.[67] However, Rupert runs the risk here of conflating the control system, which is an element of a cognitive system, with the entire cognitive system. Surely the functional elements that implement the decision of the controller by activating either short-term memory or sensorimotor mechanisms distinguish between memory and perceptual functions, but that does not mean that the controller itself is not blind to whether a strategy is conventional or extended.

Rupert's final objection is that the behaviour of the control system can be redescribed in a less sensational way. Rupert suggests that the control system selects not between conventional and extended strategies of retrieval, but between distinct *internal* information stores—in the representation constructed either from what is retrieved from short term memory, or from the visual buffer. This is quite a subtle objection and Rupert's best chance of undermining HEC lies here, I think. Certainly, Rupert's redescription of the task of the control system is not, by itself, objectionable. Nevertheless, if Rupert holds that the control system's selection of information from the visual buffer does not implicate the whole computational apparatus of the extended strategy in a functional role, he must treat the conventional strategy the same way. If the extended perceptual apparatus does not play a cognitive role in the interactive behaviour of a cognitive agent, then the conventional operations of the short-term memory apparatus cannot be considered to play cognitive roles in that behaviour either. If the environment makes a merely causal contribution to

[66] From Gray & Veksler, qtd in Clark 2008, p. 172 and re-quoted in Rupert forthcoming-a ms 32.

[67] Forthcoming-a, ms 32

interactive tasks, then the contribution of memory must also be merely causal. Some theorists may be happy to bite this bullet, but I do not see why this manoeuvre should seem attractive except for ideological reasons.

Of course, the background for Rupert's objections is much deeper than a scuffle over the significance of particular evidence. Recall that Rupert is suspicious about any attempt to apply the parity principle in advance of a theory of cognition. By offering slightly more explicit progress on the nature of cognition than Clark does, he hopes to screen off the possibility that body-external items should be called 'cognitive' even in the VCR-programming case. Rupert proposes that 'something is cognitive if and only if it is a part of a persisting, integrated cognitive system,'[68] and that putatively extended parts of cognitive realisers are not clearly parts of such systems. Rupert offers a formal scheme for identifying cognitive realisers using conditional probabilities that a mechanism or item-type will play a role in manifesting a cognitive ability, given that other mechanisms will. Mechanisms that are reliably brought to bear together for a broad range of abilities are counted as elements of a cognitive system.[69] I suggest that it is at this point that Rupert begins to beg the question against HEC, for his requirement that cognitive systems be 'persisting' precludes the possibility of soft-assembled systems. Or rather, what Rupert rules out is that the extended elements of soft-assembled systems count as cognitive components, even if they realise roles that are analogous to roles that Rupert happily calls cognitive when realised neurally. He claims that

> The persisting set of integrated cognitive capacities *is* the subject we are after; the existence of a persisting set of integrated cognitive capacities explains—partly via the positing of an architecture—why it has been empirically fruitful to proceed on the assumption that organismic subjects exercise their cognitive capacities across contexts.[70]

If Rupert is determined at the start never to call soft-assembled parts of the environment parts of a 'cognitive system' as such, it would seem that what Rupert means to indicate with the locution 'something is cognitive' is not, after all, what I mean when I say that 'something realises a cognitive functional role' (what Clark & Chalmers mean when they say that a process is cognitive). Rather,

[68] Forthcoming-a, ms 7.

[69] Cf. Rupert forthcoming-a ms 23 for a more detailed account of his proposed decision procedure.

[70] Forthcoming-a ms 37–38.

what Rupert means by the 'cognitive system' is just what Clark means by the 'persisting common biological core' of a cognitive system.[71]

It seems to me that Clark and Rupert talk past each other on this matter because they deploy alternative rhetorical strategies. Though one's choice of strategy in this matter is not entirely inert metaphysically, Clark & Chalmers and Rupert's disagreements over the application of vocabulary obscures the fact that they agree in matters metaphysical far more than they disagree. Clark & Chalmers choose their language to emphasise the surprising continuity between internal and external realisations of cognitive functional roles, whereas Rupert chooses his language to emphasise the obvious discontinuities between stable, persisting elements of cognitive architecture and those cognitive items that are realised through the opportunistic exploitation of an organism's environment. I will not criticise either rhetorical strategy, though I have adopted that of Clark & Chalmers.

Very Modest Mental Extension

So I have addressed one of the issues left from my initial discussion of Clark & Chalmers' argument. The parity rule and cognitive impartiality ground an argument for cognitive extension, where it was not clear that the Tetris example did. But what sense can be made of the mess about Clark & Chalmers' discussion of HEM?

To begin, it should be noted that HEM is both stronger and weaker than HEC. HEM is stronger in that it is more specific. Whereas HEC holds that some cognitive realisers are extended, HEM holds that a particular subset of cognitive realisers—those that correspond to rational roles like belief, desire and so on—can be extended. However, HEM is modally weaker than HEC. As I understand them, Clark & Chalmers' objective in arguing for HEM in addition to HEC is to claim that there is nothing sacrosanct about the realisers of rationally-individuated roles. In order to succeed at this dialectic manoeuvre, Clark & Chalmers need not show that cases of extended rational mentality are actual. Whereas the parity rule actually applies to e.g. the subpersonal apparatus of interactive behaviour, such as playing Tetris, programming VCRs, operating ATMs and so on, it *might* also apply to the computational apparatus subserving human rational action. Although they seem to think that it is plausible that extended mentality is actual, their dialectic aim will be satisfied so long as extended mentality is merely possible. Nevertheless, since they take HEM, like

[71] 2008, 117. Cf. Clark's discussion in 2008, pp. 116–118.

HEC, to have applications in our cognitive science, Clark & Chalmers require extended mentality to be possible for humans, and not only for some conceivable Martian.

> The Hypothesis of Extended Mentality
>
> Human mental functional roles can possibly be realised in part by entities outside of the body.

This distinction between the modal strength of HEC and of HEM explains some features of Clark & Chalmers' language that might otherwise be thought anomalous. For example, Rupert takes Clark & Chalmers to agree with his assessment of the strength of HEC, and yet is puzzled by the fact that they sometimes retreat to a modal claim. However, the retreats that Rupert cites concern the extension of beliefs, not of computationally-individuated items.[72]

There are two points I would like to make about this interpretation of HEM. First, on this view, HEM is a claim about the realisers of rational functional roles, not about rational items or roles themselves. HEM does not entail that Otto's notebook *is* a belief or a part of a belief, only that it is part of the realisation base of certain of Otto's beliefs. This also means that the truth of HEM is not closely tied to the solution to the metaphysical puzzle that we were left with at the end of section II; we do not need to know how computational and rational roles are related in order to judge whether the realiser of a rational role is extended.

Second, HEM and HEC do not have the same relation to the parity rule. HEM concerns the realisers of rational mental roles, whereas HEC concerns the realisers of cognitive roles. The parity rule concerns parity about cognitive roles, but *not* abut rational mental roles. Whether the parity rule can be used to establish instances of HEM, then, *is* tied to the puzzle from section II. If rational roles are identical to computational roles, then HEM is just a special case of HEC. Alternatively, if rational roles are realised by computational roles but distinct from them, then although the parity rule can be used to judge whether the right kind of parity obtains about rational realisers, it cannot be applied to rational functional roles. This is because the parity rule may be applicable to computationalist functional roles and yet be inapplicable to rationalist functional roles. Finally, if rational roles merely sometimes correspond to computational roles but are otherwise unrelated, the parity rule sheds no light on whether HEM is true.

---

[72] Rupert 2004, 392n7

If the relation between the parity principle and HEM is so unclear, it is understandable that Clark & Chalmers might accept the consequences of arguments from parity about computational functional roles, but not be prepared to accept all the consequences of such arguments about rational functional roles. This interpretation redeems Clark & Chalmers' equivocation about possible instances of HEM aside from the Otto case. The Otto case, after all, resembles the VCR programming case and is relatively unobjectionable on those grounds; Otto's use of his notebook involves sustained interactions with body-external information that stands in for memory, and Otto's access to the information in the notebook seems necessary for the realisation of his belief. However, Clark & Chalmers may be unsure whether to describe other cases as instances of mental extension, since the parity principle (like the parity rule) may not be a reliable guide for making judgments about rational mental roles.

This interpretation also explains the introduction of the extra conditions H1–H4 for HEM but not for HEC. Clark & Chalmers suggest that the first three conditions approximate requirements on what it means to be a belief[73]: the content of the notebook (H1) always informs Otto's actions and (H3) is endorsed by Otto. These requirements certainly have the flavour of rational requirements; the notions of endorsement, and of action (as distinct from behaviour), are rather more at home in rationalist accounts than computational ones. The requirements that Otto (H2) need not take pains to access the information and that Otto (H4) have previously endorsed the contents of the notebook may serve to preserve a rationally significant distinction between antecedent belief and learning. If the justification for all these conditions is that they approximate the contribution of a suitable account of rational belief, it is understandable why they would apply to rationally-individuated items and not computationally-individuated items. And since Clark & Chalmers are not endorsing an actual theory of belief, it is also understandable why they would not argue that H1–H4 are necessary or sufficient conditions for extension, even mental extension. They are merely factors that seem to make the Otto thought experiment easier to accept than some others.

I may seem like I am going out on a limb for HEM, but I do not mean to argue here for HEM. I am merely suggesting an interpretation of Clark & Chalmers that explains the dialectic oddness of their discussion. What I intend to gain from this suggestion is not finally for the sake of justifying HEM, but for the sake of justifying HEC. It is dialectically inappropriate for truth of HEC to be endangered by a collection of hand-wavy thoughts about HEM. The

[73] Clark & Chalmers 17

conditions H1–H4 have nothing to do with HEC, and the consequences of adapting the parity rule for rationalist theories are not consequences of the parity rule as such, or of HEC. Nevertheless, many critics have used just such strategies to argue against HEC…

IV. Sprevak Revisited

This returns us finally to Sprevak's challenge. Recall that Sprevak's discussion of HEC has two parts. In the first 'good news' prong, he fends off criticisms of HEC, arguing that HEC is entailed by any functionalist theory. In the second 'bad news' prong, Sprevak argues that 'modest' HEC contradicts functionalism, and the only stable functionalist view on cognitive extension is a 'radical' HEC that is not so much an interesting claim as it is a *reductio*. However, Sprevak makes three crucial mistakes. First, he does not relativise the properties of functional roles to the kinds of systems they subserve. Second, he employs a method for categorising of functionalist theories that is suspect. And finally, he conflates HEC and HEM throughout his discussion, and this conflation critically undermines the second prong of his argument.

Questioning the Good News

In the first movement of his argument, Sprevak develops one of Clark's[74] replies to criticisms by Rupert[75] and Adams & Aizawa.[76] The criticism goes that externally-realised 'cognitive' items have significantly different functional descriptions than internally-realised items. For example, human mental rotation is implemented in a parallel, neurological network architecture[77] in a context of significant noise and feedback from other neural systems, whereas the rotation operation as implemented in, say, a Tetris game is implemented in a serial algorithm on a silicon-based machine in isolation from significant processing noise, &c. On the basis of these drastic functional disanalogies, goes the argument, there is no case for parity, so Clark & Chalmers' arguments fail to establish HEC.

[74] 2008, in particular chapters 5 and 6.

[75] Especially 2004

[76] Especially 2001

[77] 'Neural network architecture' here refers to the structure of biological neural networks, not to the structure of classical artificial neural networks, which are inspired by neurological processes but don't necessarily describe them precisely or accurately.

The reply is that Rupert and Adams & Aizawa have drawn functional roles that are too fine-grained. Sprevak observes, as I did in my discussion of functionalism, that not every (even reliably-occurring) feature of an item should be included in the specification of its functional role. Functionalists must include certain relations and exclude others. Sprevak acknowledges that there are many ways to categorise functionalist theories,[78] one of which is that 'varieties of functionalism contain a parameter that controls how finely or coarsely functional roles should be specified (how much should be abstracted and ignored).'[79] By identifying parities at a coarser level of grain, one might articulate a functional role that corresponds to, say, a 'visual shape-rotating process.' Since this role may be realised either by native neural processes or recruited digital resources, and expert Tetris play involves actual parity about such a role, the truth of HEC is preserved.

Of course, this reply is not quite sufficient to counter the criticism if there is no reason to resist fine-grained functional roles. However, still drawing on Clark, Sprevak articulates the 'Martian intuition' which holds that

> it is possible for creatures with mental states to exist even if such creatures have different physical and biological makeup to ourselves. An intelligent organism might have 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…[80]

On other words, as the grain of a functionalist theory is set more finely, the theory becomes increasingly 'chauvinistic,' only accounting for canonical cases of cognition, such as human neurologically-realised computation. The worry about chauvinistic theories of mind is that they fail to capture what is common between canonical cognition and cognition in other creatures—be they actual terrestrial creatures or possible Martians. Since our theory of cognition should not only countenance human cognition, the grain parameter should be at least coarse enough to admit the possibility of strange, Martian realisations of cognition.[81] But Sprevak contends that 'if the grain parameter is set *at least*

[78] Cf. Sprevak's discussion of varieties of functionalism, ms 11–13

[79] Sprevak, ms 9

[80] Sprevak, ms 6

[81] Sprevak allows that for some purposes, functionalist theories might legitimately make no pretence of countenancing the Martian intuition. For example, some psycho-functionalists may be concerned only with the details of actual human cognition, and consequently unconcerned with the mind-body problem as such. However, since such functionalisms are

coarse enough to allow for intelligent Martians, then it also allows in many cases of extended cognition.'[82] After all, Martians could count by extending fleshy head-internal tubes (parity with finger-counting),[83] or Martian brains could store information about the world in the form of static ink-marks that are read back into short-term memory when needed (parity with Otto's notebook),[84] and Martians might even have stranger ways to retrieve information.[85] Thus Sprevak concludes that

> there is no intermediate setting of the parameter that: (i) allows preservation of the Martian intuition and (ii) makes HEC come out false. A vivid way of putting it is that from a functionalist point of view, the mereological sum of us and our artefacts are actual Martians.[86]

So functionalism entails HEC.

The first objection I wish to register to this first prong of Sprevak's argument is that it ignores a requirement I articulated earlier in connection with the thesis of cognitive conservation and the parity rule. I claimed that the parities that are candidates for justifying controversial instances of HEC should be parities about the functional roles that are independently established as cognitive roles *for that cognitive system* (this was condition P2.ii). When Clark deploys Martian examples, his immediate dialectic objective is not to justify cases of HEC or HEM, but to argue for what I called the thesis of cognitive conservation. Clark uses the alien symbol-memory example to stall objections (like that of Adams & Aizawa) that certain kinds of items, like the symbols in Otto's notebook, are simply not candidates for inclusion in a cognitive economy. Just as the realisation of a (non-cognitive) edge-detection function in fingerprint-analysis software does not prevent us from calling that function cognitive in the context of a different system, such as human visual processing, bizarre realisations of functions do not prevent us from seeing those functions as cognitive in Martians. And if those functions could be cognitive relative to Martians they could conceivably be cognitive relative to us as well. The point of the Martian intuition in Clark is just that cognition can take many forms.

---

unconcerned with metaphysical applications, they cannot bear the weight of metaphysical arguments about whether cognition is extended (ms 12–13).

[82] Sprevak, ms 9

[83] Sprevak ms 10

[84] Example adapted from Clark (2005, qtd in 2008, p. 91), invoked by Sprevak in ms 9–10.

[85] Sprevak ms 10

[86] Sprevak ms 10

Fine. But if those bizarre functions were actually cognitive relative to us, it would take more showing than is accomplished with the Martian intuition. If an item realises a function *Φ* that does not have a property *C* relative to one system *S*, that fact does not entail that an item realising *Φ* relative to a different system *T* also does not have property *C*. Or: *Φ*s can be *C* in *T*, even if *Φ*s are not *C* in *S*. This was illustrated in the edge-detection example: *Φ* in the software programme was not cognitive, but *Φ* in the visual cortex is. And likewise, just that a function is cognitive in a Martian cognitive system does not entail that a similar function will be cognitive in a human cognitive system.

My second objection to Sprevak's 'good news' concerns the grain parameter in functionalism. I am not convinced that functionalist accounts of the mind must vary according to some criterion called 'grain.' Certainly, particular functionalist models, either of cognitive processes or of entire cognitive systems, may be more or less human-specific and more or less detailed, but these traits do not always run together.[87] A functionalist theory may be very specific to the particularities of human cognition, without being terribly detailed: David Milner and Melvyn Goodale's two visual streams hypothesis (TVSH),[88] which concerns the gross neurofunctional architecture of human visual cognition, specifies anatomical features distinctive of human visual processing [89] without offering many details about the fine computational structure of that processing. Or a functionalist theory may be very non-specific about implementation while being quite detailed: theories of generative grammar such as Chomsky's x-bar syntax[90] give very meticulous accounts of cognitive algorithms that, while as far as we know they are only actually instantiated in humans if at all, are quite divorced from the specifics of human neural anatomy.

Another worry about grain is that it does not necessarily distinguish between distinct functionalist theories, but between alternative descriptions within a theory. A single functionalist theory can accommodate multiple levels of specificity and detail. Returning to TVSH, one can draw attention either to the human-specific anatomical details about the anatomical separation of visual

[87] Sprevak hints at the ambiguity of 'grain' when he explains that the grain of a theory coarsens as certain details are abstracted, and others ignored (ms 9, quoted above). However, Sprevak does not discuss the possibility that progressively abstracting and ignoring causal details does not track a single property.

[88] Milner & Goodale 2006

[89] The two visual streams are common to most mammals, but its distinctive functional features in human brains are the topic of the account offered by Milner & Goodale.

[90] Chomsky 1970

processing in the dorsal and ventral streams, or, in a less chauvinistic spirit, one can draw attention to the different computational requirements of sensorimotor control and object-recognition. In emphasising the former aspect of TVSH, one paints the theory as extremely chauvinistic and intolerant even of reptilian visual cognition, to say nothing of possible Martian realisations of visual processing. In emphasising the latter aspect of TVSH, one can portray the theory as describing general features of the problem space encountered by complex organisms that happen to sample electromagnetic radiation to learn about and navigate their physical environments, and how one such species confronts these challenges. So depending on how you look at it, a single theory can be either chauvinistic, with the grain too fine to admit of anything but human cognition, or general, with grain coarse enough to admit of diverse modes of realisation. Similarly, a functionalist theory may vary internally with respect to detail. A single model might describe very particular features of a cognitive process without taking them to be essential to the process. For example, in one of the more detailed parts of their discussion, Milner & Goodale present evidence that visually-derived sensorimotor schemata can be used to guide action in the absence of vision, but that these representations decay after approximately two seconds.[91] However, TVSH may take the rapid decay or visually-derived sensorimotor representations to be a normal feature of human dorsal-stream processing without taking it to be a necessary feature of vision, or even of visuomotor coordination.

Sprevak's grain parameter is ambiguous between detail and particularity, two features of functional description that are not necessarily correlated. Furthermore, these features are properly understood as modifying functional descriptions rather than entire theories. Because of the ambiguity of grain and the description-relativity of detail and particularity, we cannot identify distinct functionalist role descriptions by simply tightening or loosening our attention to details like the mesh of a sieve, with the chauvinistic roles getting caught in the finer mesh and the liberal roles falling out the bottom. Even if we sifted functional roles by particularity, thus assuring that we isolate the chauvinistic role descriptions, it is not clear that this kind of sorting helps the case for HEC. After all, since cognitive extension turns on the body-external realisation of *human* cognitive roles, worries about chauvinism are relevant only insofar as they reinforce the thesis of cognitive conservation. Chauvinistic functional roles are only bad for HEC when they are such as to beg the question against the possibility of body-external realisations of cognitive roles that are traditionally

[91] Milner & Goodale, pp. 171–174.

thought to be realised in the head. Similarly, it is only when detailed cognitive roles beg the question against HEC that they are inappropriate from the standpoint of evaluating HEC and its consequences.

It is easy to justify the relevance of extremely detailed functional roles, since they participate in our most sophisticated accounts of cognition. An extremely detailed specification of directed visual search and visual processing, complete with saccade-planning, letter-recognition and related processes of language comprehension, has obvious theoretical value if it combines elements of our best models. Likewise a detailed specification of retrieval from short-term memory. The task of identifying principled functional roles at higher levels of abstraction, however, is less straightforward. Nevertheless I have tried to suggest one robust principle for picking out apparently disjunctive roles that can ground claims of cognitive extension, based on the intersubstitutability of roles from the perspective of the cognitive system. Even this picture, however, involves detailed models of particular types of cognitive systems, such as human cognitive systems. The kind of modest HEC that I am advocating is not better served by 'coarser-grained' functionalist models than by more fine-grained ones.

Bad News Debunked

I have not yet addressed Sprevak's dangerous claim, that there is no principled defence of a modest or circumscribed HEC. Sprevak's objection here is quite unlike those of Adams & Aizawa and Rupert mentioned earlier, which interrupted Clark & Chalmers' argument at the point of applying the parity principle to actual cases. Sprevak allows that HEC can be established, but then claims to interrupt Clark & Chalmers at the point where they apply the brakes. Sprevak contends that once cognition is allowed out of the body, it gets everywhere and makes a laughingstock of itself. It is this second thrust of Sprevak's argument that, if sound, is particularly devastating for HEC-theorists. And if HEC is entailed by computationalist functionalism as I have suggested, then the danger for functionalism has not been averted by disarming the first thrust. Sprevak's understanding of HEC, however, is burdened by several critical misinterpretations. The first of these is to conflate HEC and HEM. The first sign of trouble might be that Sprevak takes both the Tetris example and the Otto example to support HEC.[92] Although Sprevak is far from alone in taking

[92] Sprevak ms 2

both these examples to support a single claim,[93] his conflation of HEC and HEM taints his arguments to a greater extent than some other critics.

On Sprevak's reading Clark & Chalmers' HEC is moderated by three necessary conditions on extended cognitive realisers, but has three critical worries about these conditions. First, they are unjustified. Sprevak observes that 'Clark and Chalmers say nothing except that [they] make HEC more modest and plausible.' [94] Second, the conditions are not satisfied by many paradigmatically cognitive processes. In order to preserve the viability of the extra conditions, then, they must apply only to body-external cognitive realisers. This stratagem, however, leads to the third worry: the conditions contradict the parity principle, which Sprevak interprets to be something like a thesis of location-indifference,[95] because they impose different conditions on extended realisers than on internal realisers. Sprevak's worries are all convincingly articulated. The problem with Sprevak's argument here is that the three conditions are the first three of the *ad hoc* conditions H1–H4 that Clark & Chalmers invoke to argue for mental extension in Otto's case. These conditions were conditions only on HEM, and not on HEC at all.

One might worry that this misunderstanding is not a problem for Sprevak's dialectic. If H1–H4 never applied to HEC in the first place then Sprevak's arguments were unnecessary, but only because he gets to dismiss them for free, as it were. However, the next phase of Sprevak's argument is also confused by his conflation of HEC and HEM. Sprevak contends that 'HEC unqualified by extra conditions… is wildly over-permissive in attributing mental states.'[96] 'Radical' HEC entails that 'simply by picking up a book, I come to believe everything contained in that book.'[97] Furthermore, Sprevak claims that

> 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 instantly swap our beliefs.[98]

these extensions quickly obliterate any pretheoretic notions about the mental or the cognitive so completely, that they constitute a *reductio* of any claims entail

[93] cf. Adams & Aizawa 2001, 53–55, Rupert 2004, 401–405.

[94] Sprevak ms 13

[95] Specifically, Sprevak re-brands the parity principle as the 'fair-treatment principle' (ms 3).

[96] Sprevak ms 16

[97] Sprevak ms 16

[98] Sprevak ms 17–18

them.[99] Here again, however, Sprevak's choice of examples reveals his conflation of HEC and HEM. The putatively extended functional roles in these examples are all beliefs, which are *rational* mental items, not cognitive items as such.

In order to establish that rational items like beliefs are extended by the parity rule, however, Sprevak would have to identify the cognitive processes associated with those beliefs, locate them in part outside of the head, and finally demonstrate the intersubstitutability of the extended computational role with a head-internal computational role relative to the system they inhabit. Sprevak does not argue this way. His argument draws instead on the Martian intuition again. Sprevak imagines that Martians could encode beliefs with head-internal ink-marks, and that Martians may even have innate 'beliefs' if they are born with some such ink-marks. The Martian may come to be aware of these 'beliefs' if it is 'sufficiently diligent' in examining its 'beliefs,' just as Sprevak may come to be aware of his 'beliefs' in the unopened book if he is similarly diligent.[100] Sprevak's strategy for identifying instances of HEC seems to be to examine cases of tool use, such as using a calculator or a book (and fringe cases of tool use like merely holding a book), and to imagine a Martian cognitive system that incorporates the gross functional features of the case. Whenever this is possible, which is most of the time, Sprevak takes cognition to have been extended in the original case of tool use.

But this use of the Martian intuition is objectionable, just as it was objectionable in his earlier argument that functionalism entails HEC. Sprevak's second major misunderstanding in the second thrust of his argument is to assume that whether a functional role has a particular property—such as being a cognitive role—cannot be relativised to the type of system. This assumption violates my condition (P2.ii) on the parity rule, which stated that in order to support the right kind of parity, a functional role should *already be accepted* as a cognitive role relative to the system of which it is a part. What pre-established *human* cognitive role is filled by the information in the book? It is difficult even to see how information on an arbitrary page of an unfamiliar book can fill any functional role at all (this was condition (P2.i)). That information is not poised to connect up with any cognitive apparatus at all except perhaps, if the book is opened to the page, through normal visual perception. But reading a sentence for the first time is not a part of a cognitive routine that is intersubstitutable with respect to an uncontroversial, head-internal cognitive item.

[99] Sprevak ms 16–18

[100] Sprevak ms 16

Sprevak might object to my use of conditions (P2.i) and (P2.ii). These are not the conditions H1–H4 that Sprevak argues against specifically, but Sprevak is not sanguine about the prospects for supplementing location-indifference with any extra conditions, presumably including my (P2.i) and (P2.ii). Sprevak gives two reasons for his pessimism. First of all, he observes that conditions supplementing location-indifference must not contradict location-indifference. But he doubts that there are any conditions that would both prevent runaway extensions, and that don't rule out possible cognitive phenomena, including Martian cognitive phenomena. Sprevak's second reason for pessimism is that '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.'[101] Both of these reasons invoke Sprevak's problematic use of the Martian intuition. However, Sprevak's dialectic goal is to undercut the principles that I cited against his Martians. We would seem, then, to be in danger of begging the question against each other.

There are two ways for me to yet undercut Sprevak's worries, however. First, Sprevak is mistaken in the first place to think that location-indifference is sufficient to establish HEC. Rupert and Adams & Aizawa all accept location-indifference without accepting HEC. Sprevak's 'good news' argument that functionalism entails HEC in fact relied on three premises: location-indifference, the Martian intuition, and his grain-parameter schema for categorising functionalisms. While I used (P2.ii) to criticise Sprevak's use of the Martian intuition, I used no such problematic aids when raising worries about the integrity of Sprevak's grain-based schema. If Sprevak's good news argument fails, the version of HEC that is entailed by functionalism is my modest HEC. The second reason not to worry about Sprevak is that the conditions (P2.i) and (P2.ii) are not merely adjuncts to location-indifference, invoked to temper it. My modest argument for HEC in section III required these extra conditions in order to establish the first instance of extra-cranial cognitive realisation without begging questions. That is, they were part of the argument that HEC was plausible in the first place, and therefore not *ad hoc*. Stripping conditions from modest HEC would not entail radical HEC; it would cripple the argument for HEC.

[101] Sprevak ms 18

V. Conclusion

Sprevak proposed a very creative form of objection to the hypothesis of extended cognition. Most critics of HEC question Clark & Chalmers' parity principle, or their application of the principle to actual cases. Sprevak forwent both of these strategies, arguing instead that moderate variants of HEC are unprincipled, and that the consequences of cognitive extension are unacceptable. Sprevak's criticisms of HEC were shown to rely on several critical misunderstandings, however. Throughout his discussion, Sprevak invokes an argument from possible Martian forms of cognition that fails to respect that different kinds of functional systems may admit of different functional compatibilities. Sprevak's 'good news' argument made unwarranted assumptions about how functionalist theories can be classified. The keystone of Sprevak's 'bad news' argument, however, the *reductio*, involves the conflation of HEC and HEM. The unpalatable consequences that are the climax in his discussion involve the critical misapplication of Clark & Chalmers' arguments. This misunderstanding, however, is not unique to Sprevak. Adams & Aizawa also conflate arguments for HEC and HEM. Rupert claims explicitly to observe the distinction between HEC and HEM[102] and often does, but then misjudges the differences in scope and strength HEC and HEM,[103] and finally argues for a weaker variation on Sprevak's *reductio*.[104] HEC is a strange claim to begin with, and interpretive foibles such as these only serve to further muddy the rhetorical waters.

My aim in this paper was to offer a somewhat refreshed perspective on the Hypothesis of Extended Cognition. I distinguished between computational and rational functionalist theories, which are subserved by different criteria for individuating functional roles. With this distinction in mind, and a sketch of the basic structure of functionalist theories, I set out to clarify the case for a modest version of HEC. I presented a refurbished version of Clark & Chalmers' original argument for HEC. I then reviewed studies by Wayne Gray et al. that provide plausible evidence that the right kinds of parity do, in fact, obtain. Finally, I revealed the misinterpretations of HEC that underlay Sprevak's arguments, and disarmed some of his more barbed claims.

I endeavoured in this discussion to preserve a relatively open-minded metaphysical perspective, and in particular I was concerned not to trivialise the distinction between computationalist theories of cognition and rationalist accounts of mentality. One consequence of this effort was that my attempts to

[102] Rupert 2004, 391

[103] Rupert 2004, 392n7

[104] Rupert 2004, 401–405

illuminate the case for cognitive extension left the case for mental extension rather obscure. But this was expected. The dialectic about HEC has got its complexities, but the relation between computation and rationality is one of the more inscrutable faces of the mind-body problem.

References

Adams, Fred & Ken Aizawa. (2001). 'The Bounds of Cognition.' *Philosophical Psychology* 14: 43–64.

———. (in press). 'Defending the Bounds of Cognition,' in Richard Menary (Ed.), *The Extended Mind* (Aldershot, UK: Ashgate).

Ballard, Dana H., Mary M. Hayhoe, & Jeff B. Pelz. (1995). 'Memory Representations in Natural Tasks.' *Journal of Cognitive Neuroscience* 7: 66–80.

Ballard, Dana H., Mary M. Hayhoe, Polly K. Pook, & Rajesh P.N. Rao. (1997). 'Deictic Codes for the Embodiment of Cognition.' *Behavioral and Brain Sceinces* 20: 723–742.

Block, Ned. (forthcoming). 'Functional Reduction,' in David Sosa, Terence Horgan and Marcelo Sabatés (Eds.), *Supervenience in Mind: A Festschrift for Jaegwon Kim* (Cambridge, MA: MIT Press). Accessed 24, April, 2009. <http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Kimfestschrift.pdf>

Chomsky, Noam (1970). 'Remarks on Nominalization,' in Roderick A. Jacobs & Peter S. Rosenbaum (Eds.), *Readings in English Transformational Grammar* (Waltham, MA: Ginn), pp. 184–221.

Churchland, Paul M. (1981). 'Eliminative Materialism and the Propositional Attitudes.' *The Journal of Philosophy* 78: 67–90.

Clark, Andy. (2001). Mindware: an Introduction to the Philosophy of Cognitive Science. New York: Oxford University Press.

———. (2005). 'Intrinsic content, active memory, and the extended mind.' *Analysis* 65: 1–11.

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

Clark, Andy & David Chalmers. (1998). 'The Extended Mind.' *Analysis* 58: 7–19.

Cohen, Jonathan L. (1989). 'Belief and Acceptance.' *Mind* 98: 367–389.

Davidson, Donald. (1963). 'Actions, Reasons, and Causes.' eprinted in *Essays on Actions and Events*, 2nd ed. (Oxford: Clarendon, 2001), pp. 3–19.

———. (1970). 'Mental Events.' Reprinted in *Essays on Actions and Events*, 2nd ed. (Oxford: Clarendon, 2001), pp. 207–225.

———. (1974). 'On the Very Idea of a Conceptual Scheme.' Reprinted in *Inquiries into Truth and Interpretation*, 2nd ed. (Oxford: Clarendon, 2001), pp. 183–198.

Dretske, Fred. (1986). 'Misrepresentation.' In Radu J. Bogdan (Ed.), *Belief: Form, Content and Function* (Oxford: Oxford University Press), pp. 17–36.

Funkhouser, Eric T. (2007). 'Multiple Realizability.' *Philosophy Compass* 2: 303–315.

Gray, Wayne D. & Wai-Tat Fu. (2004). 'Soft Constraints in Interactive Behavior: the Case of Ignoring Perfect Knowledge in-the-World for Imperfect Knowledge in-the-Head.' *Cognitive Science* 28: 359–382.

Gray, Wayne D., Chris R. Sims, Wai-Tat Fu & Michael J. Schoelles. (2006). 'The Soft Constraints Hypothesis: A Rational Analysis Approach to Resource Allocation for Interactive Behavior.' *Psychological Review* 113: 461–482.

Haugeland, John. (1998). 'Mind Embodied and Embedded,' in *Having Thought: Essays in the Metaphysics of Mind* (Cambridge, MA: Harvard University Press), pp. 207–237.

Hutchins, Edwin. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Kirsh, David & Paul Maglio. (1994). 'On Distinguishing Epistemic from Pragmatic Action.' *Cognitive Science* 18: 513–549.

Lewis, David. (1980). 'Mad pain and Martian pain,' in Ned Block (Ed.), *Readings in the Philosophy of Psychology*, Vol. I (Cambridge, MA: Harvard University Press), pp. 216–222.

McLaughlin, Brian P. (2006). 'Is Role-Functionalism Committed to Epiphenomenalism?' *Journal of* Consciousness Studies, 13 (1–2): 39–66.

Milner, A. David & Melvyn A. Goodale. (2006). *The Visual Brain in Action*. 2nd ed. Oxford: Oxford University Press.

Noë, Alva. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Rupert, Robert D. (2004). 'Challenges to the Hypothesis of Extended Cognition.' *The Journal of Philosophy* 101: 389–428.

———. (2007). 'Realization, Completers, and *Ceteris Paribus* Laws in Psychology.' *British Journal for the Philosophy of Science* 58: 1–11.

———. (forthcoming-a). 'Keeping HEC in CHEC: On the Priority of Cognitive Systems.' Page references are to manuscript pages 1–38. Accessed 28 March, 2009. <http://spot.colorado.edu/~rupertr/HECinCHEC.pdf>

———. (forthcoming-b). 'Systems, Functions, and Intrinsic Natures: On Adams and Aizawa's *The Bounds of Cognition*.' *Philosophical Psychology*. Accessed 30 July, 2009. <http://spot.colorado.edu/~rupertr/Adams_Aiz_Review_Rupert.pdf>

Schick, Frederic. (1991). *Understanding Action: an Essay on Reasons*. Cambridge: Cambridge University Press.

Searle, John. (1980). 'Minds, Brains and Programs.' *Behavioral and Brain Sciences* 3: 417–457.

Sellars, Wilfrid. (1962). 'Philosophy and the Scientific Image of Man,' in Robert Colodny (Ed.), *Frontiers of Science and Philosophy* (Pittsburgh: Pittsburgh Press). Reprinted in Wilfrid Sellars (1963), *Science, Perception and Reality* (London: Routledge & Kegan Paul), pp. 1–40.

Simon, Herbert. (1982). *Models of Bounded Rationality*. Vols. I & II. Cambridge, MA: MIT Press.

Sprevak, Mark. (forthcoming). 'Extended Cognition and Functionalism.' *The Journal of Philosophy*. Page references are to manuscript pages 1–27. Accessed 06 May, 2009. <http://people.pwf.cam.ac.uk/mds26/files/Sprevak---Extended%20Cognition.pdf>

Wilson, Robert A. (2001). 'Two Views of Realization.' *Philosophical Studies* 104: 1–31.