Philos Stud
DOI 10.1007/s11098-010-9600-6

Cognitive systems and the supersized mind

Robert D. Rupert

In *Supersizing the mind: Embodiment, action, and cognitive extension* (Clark 2008), Andy Clark bolsters his case for the extended mind thesis and casts a critical eye on some related views for which he has less enthusiasm. To these ends, the book canvasses a wide range of empirical results concerning the subtle manner in which the human organism and its environment interact in the production of intelligent behavior. This fascinating research notwithstanding, *Supersizing* does little to assuage my skepticism about the hypotheses of extended cognition and extended mind. In particular, *Supersizing* fails to make the case for the extended view as a revolutionary thesis in the theoretical foundations of cognitive science.

1 Clark's case for extension

The primary theme of Chapter 1 represents one of the book's most important conceptual threads: the idea of information self-structuring. Here is one version of the thesis, having particularly to do with perceptual information:

> The embodied agent is empowered to use active sensing and perceptual coupling in ways that simplify neural problem solving by making the most of environmental opportunities and information freely available in the optic array. (p. 17)[1]

This sort of active sensing comes in a variety of forms, but two aspects of it are central to Clark's presentation: (1) that the cognitive system learns more efficiently

[1] All page references are to Clark (2008) unless otherwise noted.

R. D. Rupert (✉)
Department of Philosophy, University of Colorado at Boulder,
Campus Box 232, Boulder, CO 80309-0232, USA
e-mail: robert.rupert@colorado.edu

Published online: 26 August 2010

by detecting correlations between its self-generated movement and the resulting perceptual or kinesthetic signals and (2) that the agent intentionally moves so as to try to produce data that exhibit such correlations.

I see little connection here to the extended view—the view that human cognition literally comprises states, property instances, or processes beyond the boundary of the organism. The correlations in question hold *between structures within the organism*; in Clark's examples, the events that constitute learning all amount to the recording of correlated patterns of activity within the organism or, in cases of AI, within a neatly bounded artificial system. Surely external material plays a historical role in producing those traces (cf. Rupert 1998), but Clark does not take the extended view to be a thesis about the subject's history of causal interaction with the environment (p. xxvii). What, though, is the role of external material as it contributes to learning via informational self-structuring, if not historical?

Rupert (2004) distinguishes between HEC—the hypothesis of extended cognition—and HEMC—the hypothesis of embedded cognition. The former is the extended view as described above. The latter, HEMC, holds that the human cognitive system is organismically bounded but that it interacts to a surprising extent with external materials in the course of its cognitive processing. While reading *Supersizing*, I repeatedly found myself thinking that Clark had provided clear examples of HEMC-based, but not HEC-based, cognitive processing. Here is Clark, quoting Lungarella and Sporns: "the agent's control architecture (e.g. nervous system) attends to and processes streams of sensory stimulation, and ultimately generates sequences of motor actions which in turn guide the further production and selection of sensory information" (p. 17). The control architecture issues motor commands and, as a result, indirectly produces sensory stimulation—and the commands, the stimulation, and the resulting correlations between them are all internal. Clark goes on to describe research by Fitzpatrick and Arsenio that involves "the cross-modal binding of incoming signals" (p. 18); but these are incoming signals in the standard sense: they enter into a robot's computational system through peripheral sensory channels (or are produced internally via proprioception). Over the following pages (pp. 19–21), this theme recurs in a handful of further examples, always to the same effect. A similar diagnosis applies to the later discussion of sensorimotor contingencies (p. 23) (as well as the discussion of sensory surrogates [pp. 35–36]). What is it to learn such contingencies? It is to have the physical materials of one's body, mostly one's brain, altered in certain respects. This is clearly an internalist view, HEMC, not HEC.

Chapter 2 introduces the idea of a 'negotiable body': under certain conditions, the brain incorporates external elements into the body schema, treating these as part of the subject's own body. For instance, neurons in macaques trained to retrieve food using rakes take on new receptive fields, suggesting that trained macaques' brains treat the rakes as extensions of the monkeys' own hands (p. 38). Prior to training, certain bimodal neurons are distinctively sensitive both to touch on a particular area of the hand and to visual stimulus of an object approaching that same part of the hand. After training, these neurons are specially sensitive to visual stimulus of objects in the vicinity of the rake head, in the way they previously had been to visually presented objects near the relevant portion of the hand.

In these cases, the cognitive story seems to me to be wholly nonextended; in fact, this seems to follow from the very nature of the evidence at issue. Research on neurons in macaques' intraparietal sulcus may show that macaques represent their bodily boundaries differently after being trained to collect food with a rake, but to the extent that the research shows this, it does so by showing *that macaques use neural resources to represent their bodies in a new way*; and neural resources are, of course, inside the organism. Internal, neural resources represent bodily boundaries, track ongoing activity of the body, and send motor commands to "body" parts, whether or not the parts so commanded are components of the organism.

To be fair, Chapter 2 contains intimations of at least two further arguments, one phenomenological, the other broadly evolutionary. I leave discussion of these mostly to other venues (see Rupert 2009b, Chapters 7 and 8; Rupert 2009c). One version of the evolutionary argument focuses on environmental tailoring or suitedness and is particularly related to results in cognitive science; so, I say a bit about it here. This argument appeals to the role of representational resources: "[T]he effect of extended problem-solving practice may often be to install a kind of motor-informational tuning such that repeated calls to epistemic actions become built into the very heart of many of our daily cognitive routines. Such calls do not then depend on…*representing* the fact that such and such information is available by such and such a motor act" (p. 75). The idea seems to be that, if a fact about the world is not explicitly represented, yet some cognitive process functions properly only when that fact holds, then the part of the world constituting that fact becomes a literal part of the cognitive process.

This is curious style of argument, resting as it does on one of the central insights of the embedded view: that certain heuristics employed by the local computational (or connectionist, or dynamical) system are valid only when employed in an environment of a certain sort (McClamrock 1995; Gigerenzer 2000). Moreover, it seems quite sensible to say that the cognitive system adjusts—either developmentally or evolutionarily—to its environment. This, however, presupposes the existence of a cognitive system that is becoming so suited. To take the tailoring process to bring into existence a further cognitive system serves no purpose. Compare: As one climbs a very high mountain, one's breathing adjusts to the changes in atmospheric pressure and density, but this provides no reason to introduce a new biological unit, the organism-plus-atmospheric-pressure-and-density. Otherwise indispensable theoretical constructs—the organism, its properties, and the ways in which they interact with environmental factors—do all of the necessary explanatory work.

Another theme touched on briefly in Chapter 2 is that of transformation: the appearance of "novel properties of the new systemic wholes" (p. 33) at work in extended cognitive processing. Chapter 3 explores this idea to a much greater extent, with regard to the transformational contribution of external codes (that is, public languages and other systems of external symbols, such as mathematical symbols—pp. 50–53). Clark argues that these material symbols transform human cognition (pp. 50, 57), conferring upon humans a wide range of capacities distinctive of human intelligence. It is, for example, only by being able to represent our own thoughts that we humans become able to think about our own thoughts, an

ability at the root of many of our impressive cognitive achievements (p. 58); and on Clark's view, we become able to represent our own thoughts only because an external code is available.

This observation does not seem to support HEC. The contributions in question appear to ground only a historical, causal account of the effects of external codes on cognition. An entirely orthodox view is in the offing, then: elements in the external code cause the activation of various mental representations, including representations of external sounds and inscriptions; these internal representations participate in internal cognitive processing.

Why should Clark object to this relatively mundane, internalist view? After all, Clark asserts that, in the important case of number words, "there is (at least) an internal representation of the numeral, of the word form, and of the phonetics" (p. 52). This, however, recognizes the essential representational materials posited by a typical internalist approach. Clark's objections to the internalist story seem to be that internal representations of words are "shallow, imagistic inner encodings" (p. 238; p. 53) and not, individually, "*fully content-providing*" (p. 52). It is not clear, however, in what way this conflicts with the internalist standpoint. Consider, for example, that computational models commonly incorporate pointers (Newell and Simon 1997/1976), which seem about as shallow as mental representations get; thus, the shallowness of mental representations of external symbols does not conflict with orthodox approaches in cognitive science. Neither does the imagistic nature of representations of public symbols. Computational primitives need not take any particular form, so long as they're treated as primitives by the computational system. Thus, there is no reason a computational primitive cannot possess pictorial or imagistic properties. So long as the imagistic properties play no role in cognitive processing, then a computational account of that process remains as viable as ever.

But, what if the particular form—the physical implementation or realizer—of a given mental representation (individuated in terms of its content) varies from subject to subject (say, from the speaker of one language to the next)? That is, what if two subjects form substantially different shallow, imagistic representations of number words with the same content (both referring, for instance, to 98)? Won't the imagistic features of the representations govern the subjects' responses in at least some circumstances? Perhaps, but that shows only that computationalism leaves something out, not that there is anything extended about the story. It is one thing to say that certain behavioral variables are distinctively affected by a vehicle's imagistic properties; it is quite another to hold that the vehicle itself is external. In the standard language-based case, the vehicle with imagistic properties is still an *internal* vehicle.

With regard to something's being "fully content-providing," the reader should ask for clarification. Does Clark think that every genuine Mentalese symbol must enter into all of the internal relations that might be relevant to *any* processing concerning what we might take to be represented by that symbol? That the mind contains modules, computing in a proprietary code, has been a highly influential view in orthodox cognitive science (Fodor 1983). It is virtually guaranteed that in any such architecture there will be at least two distinct symbols (that is, mental representations over which computations are performed) with the same referent;

moreover, it is virtually guaranteed that neither of these symbols is fully content-providing, simply because, by the nature of the architecture, one of the symbols (say, inside the module) enters into computational processes that the other symbol (in central processing, say, or in a different module) doesn't. Given this, it is no departure from orthodox, internalist cognitive science to introduce mental representations that fail to be fully content-providing.

Clark is impressed also by the way in which external symbols can, when immediately present, seem to play an active, attention-directing role in cognition (pp. 48, 57). I'm inclined to think words do play such a role, but that they do it via the activation of internal representations. Consider a recurring example drawn from the work of Dana Ballard and his associates (Ballard et al. 1997). Subjects are shown a pattern of colored blocks—the target—and are given various colored blocks as resources to use to replicate the target. Ballard et al. showed that subjects often (but nothing close to exclusively) use a strategy that relies more on looking back and forth than it does on the committing of lots of information about the target to internal memory.

We should not, however, misinterpret these results. The experiments do not show that subjects don't rely on mental representations of block colors or positions. To the contrary, one of the commonly used strategies (the P–D strategy—Ballard et al. 1997, p. 732) relies heavily on internal memory. Moreover, even on the least memory-intensive strategy—the one that involves the most looking back and forth—the deictic pointers used by subjects must represent the colors of the external blocks or their positions, even if only one block and one property at a time. What's interesting about visual pointers is the dynamic reassignment of them to the job of representing various external things, positions, or colors. Each time one is "reassigned," however, it must be bound to standing representations of properties, or else it is useless in the copying task. Comparing two bare pointers to each other or comparing one bare pointer (aimed, for instance, at the color of a block the subject has just attended to) to the color of a block in the resource pool does not do the subject any good. The subject must be able to "decide" whether the pointer and the visual representation of the color of the block to which she is currently attending (while looking at a candidate block in the resource pool) are the same, so that she can pick up the correct block. This requires binding the pointer to an external object but also to an internal representation of its color. After all, a bare pointer has no content, so the use of it alone would not guide the subject to pick up a block of one color, rather than a different one, from the resource pool. Ballard et al. do not deny this; rather, it's built into their approach (Ballard et al. (1997, p. 725).

Return now to the case of words. When reading, some words differentially capture the subject's attention. Nevertheless, it's reasonably clear that mental representations of words commonly contribute to cognitive processing in the absence of the actual units: during literature exams, students routinely produces names of characters and descriptions of settings, without having the text at hand. So, there is independent reason to posit internal mental representations activated in subjects while reading. In which case, the attention-directing role of external resources begins to look pretty humdrum: when one looks at a given word, it "directs one's attention" by causing the activation of an internal representation of that word.

2 Cognitive systems

In the second of *Supersizing the Mind*'s three major divisions, Clark responds to critics. Chapter 6, in particular, provides a sustained rejoinder to my concerns about the competition between HEMC and HEC. Some of Clark's remarks in this regard seem misleading—a matter of responding to arguments I have not propounded—see Rupert (2009c) for a detailed defense of this claim and an attempt to straighten out the dialectic.

Let me focus here on a more positive project. In previous work (2004, 2009a, b) I argue that the debate over extended cognition largely boils down to the question of how properly to individuate cognitive systems. On the view I propose, something is cognitive if and only if it is the state of a cognitive system, where a cognitive system is the persisting collection of mechanisms the integrated functioning of which causally explains, case-by-case, instances of intelligent behavior. Cognitive processing is not simply the activity of whatever causally contributes to the production of intelligent behavior. Rather, the genuinely cognitive processes are the activities of the fundamental explanatory construct of cognitive science, the cognitive architecture (which Wilson 2002, p. 630 calls the 'obligate system').[2]

How does Clark respond? As Clark sees things, the HEMC-cum-systems-based approach elevates "anatomic and metabolic boundaries into make-or-break cognitive ones" (p. 138); but it does no such thing, at least not if "make-or-break" implies that the barrier is absolute or that some interest *in the barrier itself* drives the arguments in favor of HEMC. The arguments for the HEMC-cum-systems-based approach rest on (1) the privileged causal-explanatory role of the persisting integrated architecture, (2) longstanding and successful uses of the construct of a persisting architecture that interacts with various resources in its environment to produce behavior, and (3) the superfluous nature of a HEC-based redescription of this research strategy. These arguments arrive at a nonextended conclusion from contingent facts about past successes and the application of methodological principles such as simplicity and conservatism.

Notice, too, that the HEMC-cum-systems-based view depends in no way on there being a Cartesian Theatre, in contrast to Clark's suggestion that HEMC depicts "outer resources as doing their work only by parading structure and information in front of some thoughtful inner overseer" (p. 137). In "The extended mind," Clark and Chalmers (1998, p. 17) tentatively suggest that internal consciousness must validate the cognitive status of external states. In Rupert (2004, pp. 404–405), I argued that such a view runs toward HEMC more than it does HEC. We must, however, keep the logic straight here. It is one thing to assume, as I did, that if there is a privileged internal consciousness before which structure and information must be paraded in order that they be cognitive, then HEMC (most likely) wins the day. It is quite another to assume that, if HEMC is true, there is a privileged internal

[2] Rupert (2009b) proposes a formal measure of systemic integration, one that measures the degree of one sort of interdependence among various mechanisms—internal or external—that produce intelligent behavior. As a provocative side note, I wonder whether Clark, in his discussion of the quantification of embodiment (p. 215), has hit on better measures, but ones that are still likely to yield internalist results.

consciousness before which structure and information must be paraded in order that they be cognitive. My criticism of Clark and Chalmers's argument for HEC (about dispositional beliefs) in no way presupposes the second conditional, which I take to be false. My arguments for HEMC require that the architecture be inside the organism, but they do not preclude the architecture's being distributed. Thus, Clark's charges of "magic dust" (p. 136) mongering miss the mark.

Clark's emphasis on auto-stimulation grounds a different sort of response to systems-based concerns. Clark claims that the organismically local cognitive systems temporarily dock up to external resources to create feedback loops, or other kinds of cognition-changing cycles of ongoing activity. A real-life example involves an artist's use of a sketchpad. The artist sketches out an initial idea, and the result causes the artist to envision refinements or other changes. This process iterates, eventuating in a product that the artist almost certainly would not have created had she tried to plan out the entire sketch internally, prior to execution. When cognition relies on self-stimulating loops, the cognitive system itself encompasses the entire loop, and thus to the extent that there is a relatively persisting cognitive system, it expands during loop-involving activity then shrinks back down.

Elsewhere I have discussed the case of the sketchpad in some detail (Rupert 2009a, pp. 101–102). Clark is not convinced by my treatment, however (pp. 162–163). He points out that the internal cognitive system is itself made up of a collection of mechanisms that can be selectively impaired and some of which are asymmetrically dependent on others for their contribution to cognition; this is meant to show that the sketchpad, if second-rate in any way, is no more second rate than some of the internal mechanisms the activity of which is clearly cognitive.

It seems to me that Clark does not fully engage with the systems-based approach. The systems-based approach allows the internal mechanisms to be a grab-bag bearing a variety of different relations to each other, so long as they constitute a relatively persisting, integrated system. The persisting architecture contains mechanisms that interact so as to allow the organism to learn to use a sketchpad, to take up a sketchpad when it so desires, to create new sketchpads in the absence of local ones, and so on. This is generally true of loop-involving processes.

In response, Clark might suggest that we think systems as growing and shrinking. This view seems metaphysically profligate, however (for further discussion, see Rupert 2009b, Sect. 3.4.3). The relatively persisting set of integrated mechanisms—the architecture—has proven to be of great causal-explanatory value in cognitive science, regardless of modeling orientation (connection, computationalist, or dynamicist). Moreover, of fundamental explanatory importance is the way in which the architecture governs interactions with the environment; a further explanatory function of architectural principles is to explain more or less permanent changes to the integrated system (learning being the most obvious example). So, the standard view is committed to an architecture, a world of external objects some of which causally contribute to the production of intelligent behavior, and a story about how things in the two preceding categories interact. The approach according to which the system grows and shrinks must, if it is to match the explanatory power of the orthodox view, *posit all of the same theoretical materials and processes*; but the advocate for growing-and-shrinking adds that, during interaction, the system itself

grows to include the external material—without showing that this commitment to a growing and shrinking system is anything more than relabeling (or an analytically equivalent reparsing of the standard story).

The preceding comments reflect a general worry behind many of my concerns about HEC: that it is of no importance in cognitive science to call external material 'cognitive'. In response, Clark appeals to the work of Gray and his collaborators (pp. 118–122). In a series of experiments, Gray and associates (Gray et al. 2006) measure subjects' tendency to "choose" between the use of internal memory and the accessing of information encoded in external structures. Gray does so by manipulating the relative time–cost of the use of internally and externally encoded information. The results manifest regular relations: increase the cost of access to environmentally encoded information, and subjects are more inclined to use internally encoded information; keep the cost of external information low, and subjects use it. The cognitive system seems to "care" about only the time–cost of access to information, not about its location per se.

Clark takes this to show that the external locations are part of the cognitive system. Why, though, should we not take Gray's results to show something different: that, when there is no great cost in terms of time, the cognitive system uses resources *beyond* its boundary? Clark seems to need the following premise: a system that uses resources beyond its boundary must (or at least is very likely to) treat the external nature of the location of those resources as intrinsically relevant to the decision whether to use those resources. Otherwise, why would it matter that the system Gray discusses *doesn't* treat this difference as of intrinsic import?

Take a system with any boundaries you like. There will be almost certainly be cases in which the system accesses information outside the boundary of the system and does not treat this difference as anything more than a difference in accessibility (or time to completion, or amount of pain caused in the body to get the information, or whatever). The fact that the system fails to treat the external nature of the information as intrinsically relevant to its decisions shows, so far as I can tell, nothing about the boundary of the system; it does not show that what we might have thought was an external location is really an internal one.

Consider a further worry. I see no reason to doubt that, when the system makes use of externally encoded information, the body-bounded system forms internal representations of that information (compare this to the remarks made above about the role of representation in Ballard's results). The cognitive process of accessing an organismically external store, then, is best cast as an entirely internal process. Thus, the situation can (and should) be described in wholly HEMC-based terms: there is a competition between the use of one internal store—short-term, declarative memory, and the use of a distinct internal store—a visual buffer. Both of the locations from which information is accessed are inside the organism, and thus, the system's process of "choosing" between them appears to have no bearing on HEC.

Lastly consider a motive that seems to drive much of the interest in HEC, the epistemic role of seeing cognitive systems as extended. In earlier work (Rupert 2004), I argue that this motive provides little support for HEC. There are simply too

many cases in which understanding some phenomenon (a war, for example) requires cognizance of factors beyond the confines of that phenomenon (the military engagement itself) or that lack properties distinctive of that phenomenon (military properties). In *Supersizing*, Clark continues to advance epistemic dependence arguments (pp. 116, 157–158), for reasons that are unclear to me. If epistemic strategies are to drive metaphysical conclusions in the present case—conclusions about the location of cognition or mind—an argument for such exceptionalism is required.

To the extent that Clark offers such an argument, it rests on a pragmatic point. Clark sometimes suggests that the adoption of anything short of HEC obscures the importance of the environment from cognitive-scientific view (p. 136). There is, though, no reason to think HEMC occludes the environment's contribution to human cognition. After all, HEMC expressly encourages cognitive scientists to focus on ways in which the human cognitive system exploits external resources. Clark's concern would be more compelling were there actual cases in which the HEC-based perspective led to cognitive-scientific advances and where HEMC, had it been adopted in place of HEC, would have prevented these advances. So far as I can tell, though, the empirical research taken to support HEC was motivated not by a specific commitment to HEC or to HEMC, but rather by a general sense that interaction with the environment plays an important role in cognitive processing. (See, for example the way Ballard and colleagues describe their "central thesis"—Ballard et al. 1997, p. 723 [cf. Rupert 2004, pp. 393–394, footnote 9].)

Supersizing does very important work: it forces philosophers of mind and cognitive science to confront the messy, complex, and beautiful ways in which real human cognition proceeds. As it stands, though, this does not make a convincing case for the hypothesis of extended mind or extended cognition.

References

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*, 723–742.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gigerenzer, G. (2000). *Adaptive thinking*. Oxford: Oxford University Press.

Gray, W., Sims, C., Fu, W., & Schoelles, M. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review, 113*(3), 461–482.

McClamrock, R. (1995). *Existential Cognition: Computational minds in the world*. Chicago: University of Chicago Press.

Newell, A., & Simon, H. (1997). Computer science as empirical inquiry: Symbols and search. In Haugeland, J. (Ed.), *Mind design II: Philosophy, psychology, and artificial intelligence* (pp. 81–110). Cambridge, MA: MIT Press. Reprinted from the Communication of the Association for Computing Machinery 19 (March 1976), pp. 113–126.

Rupert, R. (1998). On the relationship between naturalistic semantics and individuation criteria for terms in a language of thought. *Synthese, 117*, 95–131.

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

Rupert, R. (2009a). Innateness and the situated mind. In P. Robbins & M. Aydede (Eds.), *Cambridge handbook of situated cognition* (pp. 96–116). Cambridge: Cambridge University Press.
Rupert, R. (2009b). *Cognitive systems and the extended mind*. New York: Oxford University Press.
Rupert, R. (2009c). Critical notice of Andy Clark's *Supersizing the Mind*. *Journal of Mind and Behavior, 30*(4), 313–330.
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*, 625–636.