The Goldilocks problem and extended cognition

Daniel A. Weiskopf

Abstract: According to the hypothesis of extended cognition (HEC), parts of the extrabodily world can constitute cognitive operations. I argue that the debate over HEC should be framed as a debate over the location and bounds of cognitive systems. The 'Goldilocks problem' is how to demarcate these systems in a way that is neither too restrictive nor too permissive. I lay out a view of systems demarcation on which cognitive systems are sets of mechanisms for producing cognitive processes that are bounded by transducers and effectors: structures that turn physical stimuli into representations, and representations into physical effects. I show how the transducer-effector view can stop the problem of uncontrolled cognitive spreading that faces HEC, and illustrate its advantages relative to other views of system individuation. Finally, I argue that demarcating systems by transducers and effectors is not question-begging in the context of a debate over HEC.

'Roland had learned to see himself, theoretically, as a crossing-place for a number of systems, all loosely connected. He had been trained to see his idea of his "self" as an illusion, to be replaced by a discontinuous machinery and electrical message-network of various desires, ideological beliefs and responses, language-forms and hormones and pheromones. Mostly he liked this. He had no desire for any strenuous Romantic self-assertion.' A. S. Byatt, *Possession*

The embodied, extended, embedded, and enactive cognition movements promise revolutionary things both for cognitive science and our ordinary conception of ourselves.[1] In cognitive science, they aim to loosen the Cartesian stranglehold on our theorizing and reorient us towards new models that recognize cognition as something not restricted to the brain, but as happening in the body and the world. Correlatively, this implies a vision of ourselves as beings whose cognitive nature is constituted in part by our bodily and worldly environments. At the extreme, the picture emerging from these allied movements depicts us as 'a vast parallel coalition of more or less influential forces, whose largely self-organizing unfolding makes us the thinking beings we are' (Clark, 2008, p. 131). This is the natural self-image to adopt if we free

[1] See the papers collected in Robbins & Aydede (2008) for a recent survey.

ourselves of the idea that in each of our brains there is a ‘Central Meaner’ whose activities most significantly constitute our abilities to reason, plan, and carry out other acts characteristic of human intelligence.

Here I will focus on the arguments for the *hypothesis of extended cognition* (HEC): the claim that some aspects of everyday cognition actually take place in the extrabodily environment. This is distinct from the *hypothesis of extended minds* (HEM), which claims that some of our everyday mental functioning (as identified by folk psychology) takes place in the extrabodily environment. Both HEC and HEM face a pointed challenge: if *some* of the extrabodily environment is part of our cognition and mentation, what is to stop vast chunks of it from also being incorporated? In short, what is the principle of demarcation that determines that this aspect of the world, but not that one, should be counted as part of the mind or cognition?

Call this the ‘Goldilocks problem’ for psychological taxonomy. The problem is to find a way of drawing boundaries around mind and cognition that is neither too wide nor too narrow, but rather ‘just right’. There may be varying notions of what counts as ‘just right’ in this debate, of course. The main criteria for an adequate solution is that it be explicit and principled. A possible further condition is that it conform with well-entrenched practices and taxonomies in cognitive science; in other words, that it be conservative. Conservatism is likely to be seen as prejudicial by HEC’s proponents, however, who aim precisely to reform these ways of thinking. But the deeper rationale for conservatism is that any criterion we advance must at least account for past successes; and ideally, if we are engaged in a revisionary project, we should also provide some sort of demonstrable explanatory advantage over the practices that underlie those successes.

I will argue that the proper locus of the debate over HEC should be how to draw the boundaries of cognitive systems, and lay out a principled demarcation criterion for such systems.[2] The proposal I defend—the *transducer-effector view*—harks back to traditional notions about classical computational systems. The upshot of this view is that most of the examples of alleged extended cognition turn out not to be. Moreover, this view can accommodate the explanatory successes of traditional cognitive science, and has significant advantages over other systems-based views in the literature. Finally, it offers a solution to the problem of cognitive spread. These constitute significant arguments in its favor.

1. The fundamentality of cognitive systems

A number of different notions have been employed in framing the debate over HEC, in particular the notions of a cognitive *vehicle*, *state*, *process*, and *system*. In this section I will attempt to clear up the relations among these various notions and suggest a plausible hierarchy of dependence among them. This will help to set the stage for the next section, where I propose a criterion for distinguishing what happens inside a particular cognitive economy from what happens outside.

HEC is sometimes described as the thesis of 'vehicle externalism' (Hurley, 1998). A vehicle of cognition is a repeatable physical structure that bears representational content. Symbols in LOT, activation patterns in connectionist networks, abstract mental models, and neural firings in various regions of cortex are all candidates for cognitive vehicles. These structures all possess physical, or more generally *formal*, properties that make them apt for

[2] Taking a systems-based approach to the demarcation problem is not novel: Rob Rupert has developed his own systems-based approach at great length (2004, 2009). However, I argue in section 5 that his own criterion of demarcation faces difficulties that can be surmounted by the approach taken here. Rupert and I agree, however, that conservatism (at least in the sense of being able to capture the successful explanations and practices of cognitive science) is an important criterion for any solution to the demarcation problem.

entering into complex causal interactions with one another, as well as for causing behavior.[3] Moreover, they are all capable of serving as representations. For my purposes, a representation is just a structure that has something like accuracy conditions, truth conditions, or satisfaction conditions, and is such that circumstances can either meet or fail to meet those conditions. Vehicle externalism, then, is the claim that at least some cognitive vehicles are located in the extrabodily world.

A cognitive *state* involves the tokening of some vehicle or other. It can sound peculiar to ask about the spatial location of a state, but if one wants to talk this way, there are two plausible proposals. First, states are located wherever the systems that possess them are; nothing more specific can be said to localize them. Second, states are located wherever their vehicles are. But in either case, states require bearers. A cognitive state is always a state of something—hence the distinction between personal and subpersonal states, which can be spelled out as the difference between states that are attributable to the whole person or whole organism and those that are attributable to mechanisms that comprise parts of the person's cognitive system. In either sense, states are never isolated. As I will understand them, cognitive states only come about in virtue of organized systems of processes and mechanisms, and they belong to the systems whose operations produce and sustain them.

Sometimes HEC is stated in terms of the spatial location of cognitive *processes* rather than vehicles. As Rowlands (2009, p. 1) puts it, it is the claim that 'at least some token cognitive processes extend into the cognizing organism's environment'. Cognitive processes are sequences of cognitive states that are produced in virtue of the operations made available by the underlying

[3] Saying what a 'formal' property is in this context is extraordinarily difficult. See Schneider (2009) for recent discussion. All that I will mean by 'formal' properties here is non-semantic properties; they may be physical, functional, etc.

architecture of the system to which they belong.[4] Different architectures, employing different kinds of representational vehicles, will have correspondingly different operations available to them. In classical symbolic systems, the operations include comparing and concatenating symbols, and transforming strings of symbols into new strings in accordance with some rule; e.g., for systems that embody propositional logic, the rules might include AND-elimination and double negation deletion. In connectionist systems the rules are those that determine how activation is passed from one layer to another and how the values of weights change over time. In systems using perceptual symbols, the rules might involve performing rotation on mental images, scanning an image for a match to a symbol, or determining the overlap in volume between two represented bodies in space. This notion of a cognitive process is generic: the operations that determine the next stage in processing can be of any sort, so long as they turn one representation (or set of representations) into another in some systematic way.

Finally, HEC is sometimes claimed to be a thesis about the spatial distribution of cognitive *systems* (Clark & Chalmers, 1998). I will say more about what systems are in the following section, but to anticipate, a system is an interlocking set of mechanisms for producing cognitive processes, along with a specification of the representations that are employed in those processes. These mechanisms are usually highly interactive and produce complex behavior only as a result of their ensemble activity. Typically, but not mandatorily, such systems are decomposable into subsystems dedicated to carrying out one kind of process or another, with

[4] On this view, not every temporal sequence of cognitive states counts as a cognitive process. This is as it should be. Our psychological lives may not be very orderly—chains of thought are usually interrupted by daydreams, pains, bouts of reminiscence, and other intrusions. Processes are started only to be superseded by others, then perhaps picked up later or abandoned altogether. To separate out these various threads in a creature's mental lives requires more than temporal ordering. It requires appeal to an *underlying* organizing principle. What a process would produce if uninterrupted is not what it produces when actually run in a creature's messy mental life.

links between the subsystems to pass information and control signals (e.g., activation or inhibition of another subsystem).

Putting these together, we get the following picture: (1) every cognitive vehicle is an element of some cognitive state; (2) every cognitive state is part of some cognitive process;[5] and (3) every cognitive process takes place in some cognitive system. This imposes conditions on what one needs to show for each variety of extended cognition.

In order to establish vehicle externalism, one would have to make the case for process externalism. A worldly representation is only a cognitive vehicle if it participates in cognitive processes. And cognitive processes only exist insofar as there are sets of mechanisms for producing and sustaining them; that is, if there are cognitive systems in which those processes play a role. So to make the case for process externalism one would have to make the case for system externalism. Vehicles, states, processes, and systems come as an interdefined package, but ultimately, no matter which approach we take, we need to say, in as neutral a way as possible, what counts as a cognitive system; or, for present purposes, we need to answer the narrower question: what makes the difference between being inside and outside such a system.

Some advocates of HEC have also come to the conclusion that the fundamental question concerns the boundaries of systems. For instance, Clark (2008) says: 'What counts are not interfaces [between organisms and the world] but systems—systems that may come into being and dissolve on many different timescales but whose operation accounts for much of the distinctive power and scope of human thought and reason' (p. 159). However, what has largely been missing from the debate—and what is necessary if we are to solve the Goldilocks

[5] Note again, this only says that every token cognitive state that is produced happens in virtue of some underlying mechanism that functions to produce states of that type. For example, the state of entertaining a mental image of a flying elephant is something that can only come about in a system that contains processes for forming such images. Put in different terms, cognitive states can only come about if there exists a mechanism that can produce not only those states but also related ones that form part of the same type of processing.

problem—is any criterion for what distinguishes the inside of a cognitive system from the outside.

This order of explanatory dependence has not been universally accepted. So Adams & Aizawa (2008, pp. 106-7) argue that it is cognitive processes that are fundamental, rather than cognitive systems. They point out that whole human beings are cognitive systems, and that one's big toe (say) is part of the whole human being, hence cognitive systems can extend to include one's toes and other parts of the body. They think that this establishes an 'informal sense' (p. 107) in which cognitive systems can extend into the body, which they take to be uncontroversial. However, the claim that cognitive *processing* takes place in one's big toe would be surprising. I agree with this latter claim, but deny that human cognitive systems include every part of the human body. The error here is in taking whole human beings to be cognitive systems. Human beings *possess* cognitive systems, but their boundaries are not those of the whole human. So if there is no cognitive processing in one's big toe, this is plausibly because the cognitive system embedded in the whole human doesn't extend to the toe itself.

Rowlands' (2009) attempt to demarcate cognitive processes also shows the need to move to an approach that takes systems as fundamental. His proposal is: a process P is a cognitive process iff (1) P involves information processing; (2) this processing has the proper function of making new information available either to the subject or to later processing operations; (3) the information processing involves the production of a representational state; and (4) the process '*belongs* to the *subject* of that *representational state*' (p. 8, emphasis in original).

The sticking point here is condition (4), the ownership condition. Rowlands rightly points out that understanding what it means for a process to be owned is an extremely difficult task, but he adds that spelling this out is a job for internalists as well as externalists. Without some such

criterion we face the problem of 'cognitive bloat' again. For instance, to borrow his example, the representations produced by my telescope as I use it to perceive Jupiter's moons would be at risk of being cognitive processes that belong to me, since they satisfy conditions (1)-(3): the telescope produces information-bearing representations, and has the proper function of doing so as a way of providing this new information to its user. Similarly, he adds, for what goes on in my computer and my calculator when I use them. All of these would count as cognitive processes if (1)-(3) were sufficient conditions for being such things, and presumably they would be *mine* (who else's?).[6]

Ownership is intended to block bloat. The admittedly tentative suggestion that Rowlands gives for spelling out the notion of ownership appeals to the integration of one process with others. Roughly, a process P is integrated with other processes Q and R 'when it is fulfilling its proper function with respect to those processes' (p. 17). In the case of cognitive processes, this presumably means that, for example, P takes its inputs from Q and feeds its outputs to R. And a process is owned by a subject iff it is sufficiently well-integrated with other processes in the subject's life. 'Ownership is to be understood in terms of the appropriate sort of integration into the life—and in particular, the psychological life—of a subject' (p. 17).

A metaphysical worry that arises with respect to this picture is that we do not yet know what a 'subject' is here. But setting this aside, this criterion seems too weak to rule out the earlier counterexamples. The telescope that I peer through *is* fulfilling its proper function of representing distant moons to its user when I use it. The telescopic processing is integrated with my own visual processing, just as its designers intended. And similarly with any other

[6] Interestingly, Rowlands treats this as a problem for HEC, although there may well be advocates of the thesis for whom they are simply natural, indeed welcome, consequences of the view. But dialectically this is fair, since both opponents of HEC and some of its defenders will want a principled way to rule out at least some cases of cognitive bloat.

extrabodily tool that I use, since tools are defined (in part) in terms of their proper functions. If integration (and hence ownership) only requires that a process be fulfilling its proper information-processing function with respect to other processes, then bloat remains unblocked.

Rowlands might try to tighten up the conditions on integration. Perhaps it's required that a process be integrated with *many* other processes for it to be genuinely owned. In section 5, we will consider Rupert's attempt to define cognitive systems in something like this way. Or perhaps the integration has to take a specific form. However, for the time being, my diagnosis is that the mistake at work here is starting with the notion of a cognitive process and then trying to spell out what it is for these processes to be integrated with a 'subject'. We can make greater progress if we start with the notion of a cognitive system, and explain what it is for a process to be taking place in that system by appeal to the demarcation criteria for such systems in general.

2. The transducer-effector view of systems demarcation

The conception of a cognitive system that I will be working with is one that derives from Pylyshyn's discussion of cognitive (functional) architectures (1984, pp. 30-1). A cognitive system is a set of physical structures and mechanisms that collectively realize a specific functional architecture. Such an architecture makes available a representational vocabulary, a set of primitive operations defined over them, a set of resources that these operations may make use of, and a set of control structures that determine how the activation and inhibition of operations and resources is orchestrated. These collectively determine the internal dynamics of processes in the system: how one set of input representations triggers a cascade of processing throughout various parts of the system, resulting eventually in some sort of output.

Within this generic definition of a cognitive system, there are many more determinate ways to fill in the details of the architecture, and much of the debate among working cognitive psychologists and neuropsychologists centers on this problem. The specific sort of architecture that is at work in human cognition is not our main concern here, however. Neither is the rather difficult question of how we are to individuate types of architecture. Rather, what is relevant is that the conception of cognitive systems as sets of mechanisms that realize a functional architecture comes with a criterion for deciding what is internal to the system and what is external to it.

The criterion is this: the boundaries of a cognitive system are given by the location of its transducers and its effectors. A *transducer*, in Pylyshyn's terms (pp. 151-178) is a device that (1) maps inputs described in physical terms into outputs described in representational terms in a way that is (2) interrupt-driven and (3) primitive and nonsymbolic. Saying that transducers are interrupt-driven is just to say that their activation is mandatorily determined by the presence of their physical input conditions. Saying that they are primitive implies that they do not carry out their mapping function by any internal representational means; their operations do not involve cognitive processes, although they may obviously be physically complex.

The most important condition on transducers, for our purposes, is that they have the function of turning physical stimuli into representational or computational states. The inputs to a transducer are not themselves representational; transducers respond only to physical properties and magnitudes. They take, for example, pressure, temperature, vibrations in the air, or ambient light in a region of space, and produce vehicles that *represent* something, most frequently some aspect of the environment that the stimulus typically carries information about. Transducers can

thus be thought of as the place in where things in the external environment become *input for* the cognitive system.

The same can be said of effectors. Corresponding to the above definition of a transducer, an effector is a device that (1) maps inputs described in representational terms into outputs described in physical terms in a way that is (2) interrupt-driven and (3) primitive and nonsymbolic. That is, an effector does what a transducer does, but in reverse. It takes a representation and produces a physical event; for example, an activation pattern in certain muscle groups. The input representation can be understood as something like a direct motor command, and this command acts immediately on the body. Both transducers and effectors are important for delimiting systems, but for brevity I will sometimes simply call this the *transducer view* of systems.

A naïve view would be that transducers and effectors are located at the periphery of the organism: Merkel cells are distributed throughout the skin, rods and cones cluster in the retina, and so on. But this is a mistake in two ways. First, there are internally located receptors that respond to various conditions of the organism. The gut, for instance, is densely innervated and can modulate brain activity in complex ways (Gershon, 1998). These interoceptive sensors may be distributed widely throughout the body. More importantly, there is a difference between sensory receptors and transducers. A single Merkel cell or rod taken on its own may not constitute a transducer. Whether something is a transducer depends on whether its output is representational. It may be that larger arrays of neurons are required to produce representations that can be used by later processing systems. A transduction mechanism, then, can itself be a large collection of interconnected neural processing units.[7] What matters is that its internal

[7] The references to neural processing units here should not be taken to imply neural chauvinism, obviously. Many systems use artificial non-neural transducers.

operations themselves are not computational or representational. This is what justifies our treating it as a *primitive* processor from the point of view of the architecture. It may be difficult to determine how to 'chunk' a complex neural system into those parts that carry out the function of transducers and effectors. The point is not the minimize these complexities, but only to note that the notion of a peripheral sensorimotor cell and the notion of a transducer-effector need not always coincide.

What does it mean to be 'within' the boundaries of transducer-effectors? Physical containment is neither necessary nor sufficient. What matters is that something take its input from them, or deliver its output to them. Normally, in the case of biological organisms, this will involve inbound or outbound spatial movement, but it need not. We can easily imagine strange creatures that have their transducers on their bodily surfaces, but keep their central nervous system elsewhere. Dennett's thought experiment in which a series of mishaps result in his ending up as a brain in a vat connected by radio signals to a distant body is a perfect example (Dennett, 1978). The physical location of further processing components is irrelevant; what matters is their functional connectivity. This point made by extended cognition theorists is surely correct. If a physically distributed system can realize a functional architecture, then cognitive systems may be widely physically distributed. So the transducer view is hardly to be stigmatized as positing the skin as a 'magic membrane' between mind and world.

The motivation for adopting the transducer view can be seen in Pylyshyn's discussion. He remarks that

> aspects of the physical environment to which the computer may be called on to respond—say, in a machine vision or speech recognition system—generally are not the same as the aspects that define computational states. Indeed, rarely is a physical event

> outside a computer's mainframe considered a computational event (though, with distributed computation, the boundaries of the "computer" become less well-defined). Computational events invariable are highly specific equivalence classes of electrical events within the machine's structure. If, however, we have a device that systematically maps interesting equivalence classes of physical events into classes of computationally relevant internal events, we encounter the possibility of coupling the machine—as computer—to a noncomputational environment. (Pylyshyn, 1984, pp. 151-2)

A virtue of this account, then, is not that it merely gives us a way of telling inside from outside. It also does the much more important job of telling us what sorts of events count as *input* to the system and *output* of the system.

It is possible to *influence* the course of processing in a system in any number of ways. A simple knock on the head may produce thoughts of being Napoleon or hallucinations of pink bears. The knock on the head is the cause, but it is not an input, since the system is not designed to produce those states in response to head-knocks (see Block, 1978). More sophisticated techniques like transcranial magnetic stimulation may selectively disrupt neural activity in certain brain regions, but those regions do not have the function of inactivating in response to magnetic fields. The inactivation arises simply from altering the realizing structures for part of the system. Only alterations that are mediated by transduction and effection count as input and output, however.

The transducer view has substantial initial plausibility. It provides a clear criterion for distinguishing cognitive systems from their environment, and in doing so helps us to make the important distinction between what is properly input to and output from these systems. Its further virtues will emerge as it is compared to its rivals.

3. Skepticism about transducers

Haugeland (1998) has argued that the notion of a transducer is fundamentally a confused one, and that focus on it distracts us from the important facts concerning how organisms interact fluidly with their environments. He offers several related arguments for the conclusion that a theory of behavior should dispense with the notion of transducers and effectors entirely. None of these, however, is persuasive.

Haugeland proposes that transducers are inherently 'low-bandwidth' devices (p. 220). That is, they take a relatively information rich stream of stimuli from the world and squash it down to a few bits encoded in a symbolic description. But this, he conjectures, results in a system that loses significant capacity to respond sensitively to the details of the perceptual situation. A system lacking this sort of 'bottleneck' could engage more fluidly with its surroundings. So we should reject the transducer conception of how cognizers relate to the world, in favor of a non-transduction based 'high-bandwidth' interaction.

But as Clark (2008, pp. 31-3) points out, it is a mistake to suppose that all transducers need to be low-bandwidth.[8] This seems to be an illusion generated by Haugeland's focus on *symbolic* descriptions as the output of transduction. Symbols, in something like the LOT sense, are one possible output, but it is equally possible that transducers output elements of fairly fine-grained perceptual models of the environment. These perceptual symbols (Barsalou, 1999) are not linguiform, and may encode robust detail about their inputs. Transducers may mediate fairly high-bandwidth interactions. Moreover, there may be low-bandwidth connections within the

[8] Strictly speaking, Clark doesn't use the language of transducers. He talks instead of interfaces, and argues for a conception of interfaces as points at which one system can be detached from another to operate on its own or to be recoupled to a different system. None of his points against Haugeland depend on adopting this conception, however.

cognitive system. A control system may send messages of only a few bits to other systems, but this may be enough to sensitively orchestrate wide-scale changes in cognitive functioning.

Haugeland also argues that the corresponding notion of an effector should be abandoned. His argument rests on the notion that the instructions sent to effectors must be syntactic expressions whose content 'does *not* depend on how or whether [they] might be acted upon by any *particular* physical output system' (p. 225). It ought in principle to be possible to plug anyone else's hands, or even an artificial robot hand, into my effectors and get the same movement as a result. But the signals that are sent to my fingers to get them to perform a specific action must take into account all sorts of facts about *my fingers*—their length, muscular capacity, flexibility, and so on.

Nothing in the notion of an effector, however, requires that it have this sort of device-independent content. If effectors only took symbol-like descriptions pitched at a high level of abstraction, this might be the case. But any sort of representation available to the system may function as an effector's input. Motor representations may encode patterns of intended movements as vectors, as motor 'images', or in any other way. The nature of the task, here as elsewhere, shapes the representational resources used to carry it out. In fact, for each individual, there may be subtly different motor codes used that are shaped by and reflect the properties of the body that they have been adapted to control. Cognitive systems may have components that cannot be separated from their normal bodily environment and plugged into a new body. All that is required is that this system be capable of being plugged into a *relevantly similar* body, meaning here one capable of executing the command coherently.[9]

[9] Haugeland's point is perhaps that there isn't and couldn't be any other body but mine that can execute the finely tuned commands my effectors receive; hence his disparaging remarks about 'God's own microsurgery' being inadequate to the task (p. 225). But this can hardly be an objection to there being body-specific commands that feed effectors.

Finally, Haugeland suggests that transducers and effectors are often theoretically idle, since there are many tasks that can be performed without the mediation of any sort of perceptual representation. Rather, many tasks, such as retrieving the milk from the back of a cluttered refrigerator, involve tightly time-locked perception-action cycles. Nothing like 'reasoning' needs to be implicated in this task (p. 221), and if reasoning is unnecessary, there is no need for transduction either, since the only point of transduction is to present inputs to reasoning processes.

This argument rests on the assumption that every cognitive process mediating perception and action needs to count as 'reasoning' in some possibly inflated sense. Quite the contrary; these process may be domain-specific, rapid, heuristic-laden, and operate only on the assumption that innumerable unstated conditions are met. Even environmentally guided search requires some sort of internal guidance, even if only in terms of representing what the target of the search is and what should be avoided in carrying it out. Sensorimotor engagements needn't involve reasoning to involve representing, and insofar as they involve the latter, they need transducers and effectors to interface with. Advertisements to the contrary notwithstanding, then, the transducer view doesn't appear to be incoherent or idle.

4. Inputs, influence, and interfaces

A virtue of the transducer view is its ability to block HEC's perennial nemesis, the problem of cognitive bloat. Proponents of HEC note the undesirability of letting any mere causal influence on cognition count as part of a cognitive system. Rowlands (1999) offers one view that I think nicely captures the idea of a cognitive process at work in much of HEC: 'The manipulation of external structures is a process which is essential to accomplishment of a

cognitive task…. Moreover, it involves operations on information-bearing structures; structures which carry information that is relevant to the task at hand. Therefore, it counts as a cognitive process' (p. 116). But as we saw earlier, allowing any information-bearing structure that helps to complete a cognitive task to be part of a cognitive process (and hence part of a cognitive system) leads inexorably to the causal spreading of cognition in undesirable ways. In daily life we are in contact with innumerable external sources of information that assist us in cognitive tasks, but not even the most ardent proponents of HEC want to count *all* of these as part of our cognitive processing. My Google queries ultimately may manipulate some data on a server stored in Helsinki, but that distant server is not part of my cognition, nor are the server and I part of any single cognitive system.

In a similar vein, Clark (2008, p. 130) says that if the noise of the rain on my window on a typical day in Edinburgh just happens to help my thoughts to flow along in productive ways, we should not count the rain as part of my cognitive system. The reason is that 'the rain is not part of… any system selected or maintained for the support of better cognizing. It is indeed *mere* (but as it happens helpful) backdrop.' Adding this selection criterion is intended to filter effects from system components. This implies that for a robot that is designed to use raindrop sounds in order to time its internal operations, the rain *would* count as part of its cognitive system. The same would go for a more sophisticated robot that was designed to produce the very external signals that aid in its own cognition through cycles of self-stimulation. Much of human cognitive activity does consist in producing external events—gestures, marks on paper, organizations of objects in space—that systematically aid us in performing tasks. So, the argument goes, we should recognize these as genuine parts of cognition, at least so long as they persist.

These examples, however, are still vulnerable to the cognitive bloat objection. For instance, suppose that I find it nearly impossible to write unless I'm in precisely the right sort of chair, drinking the right sort of coffee, and wearing shoes that keep my feet sufficiently warm. If provided with these things, my writing soars along; if not, it falls flat. (Writers have needed *much* more baroque arrangements in order to work.) Hence I arrange to have all three at hand when there's writing to be done. They are in fact selected (by me) for their role in facilitating my cognition. But intuition balks at thinking of my chair, coffee, and shoes as part of my cognitive system.

This case shows the importance of distinguishing between *influences* that are necessary for optimal performance of a task and *components* of a system. A further virtue of the transducer view is that it enables us to distinguish between those influences that are *inputs* to a system and those that occur across *interfaces* of a system.

One example of an interface is the standard peripheral slots attached to the bus of a desktop PC. These slots take expansion cards that allow various new functions—graphics display, sound, networking capabilities, etc.—to be added to the computer. While information is exchanged across via these slots, it doesn't go through a process of transduction to do so. The physical coupling between card and motherboard allows computational instructions and data to be directly transmitted between the two. Call any physical structure that allows representations to be passed back and forth in this way an *interface*.

Subcomponents and subsystems of both artificial computers and cognitive systems are generally related via interfaces. Spike patterns in the lateral geniculate nucleus may be transformed in complex ways as they are distributed to regions of V1, but those signals are never transformed into a brute, nonrepresentational signal and then re-encoded representationally by a

transducer. This is characteristic of cognitive processes generally: they are sequences of states that are produced and sustained by mechanisms that transform one representation directly into another by means of a set of primitive operations. The domain and range of these operations are defined in terms of these sets of representations: a rotation operator takes one visual representation and produces another one, a parser (considered as a unitary operator) takes an orthographic or phonological linguistic representation and produces a syntactic one, etc. Transducers are not interfaces in this sense, then, since their domain is physically characterized stimuli, rather than sets of representations.[10]

The inside of a complex cognitive system is decomposed into components that are linked by interfaces. The way to get a larger cognitive system out of two independent systems is by joining them via an interface. This is the force of Pylyshyn's point about distributed computation: where many computers are joined either by physical cables or network connections, they are interfaced, and hence comprise a larger system. Thanks to ever-present wireless networks, most of the computers that we use on a day to day basis are not really independent units at all, but part of a large system spread out over a shifting and perhaps indeterminate physical terrain.

Adopting the transducer view thus allows us to make a set of theoretical distinctions that are invisible on many versions of HEC. The distinctions are between mere influences on a system, inputs to the system, and interfaces within a system and between systems. These

[10] However, we should note that although cognitive systems are representation-processors, that doesn't entail either that everything inside of a cognitive system is (or involves) representations, or that there are no representations outside of cognitive systems. There might be 'natural representations' in the world in addition to various sorts of public, derived representations. However, none of these are *cognitive* or *mental* representations. Being within a transducer-effector delimited system is a necessary condition for this. See section 6 for further discussion of this point.

distinctions will be relevant in defusing objections to the transducer view in section 6. I turn now to contrasting the transducer view with another prominent view of how systems are individuated.

5. Rupert's systems based view

Rupert (2004, 2009) also holds that the central theoretical issue is whether cognitive systems extend into the extrabodily world. His systems-based view is an attempt to give a set of criteria for demarcating what is part of a system from what isn't. While I am sympathetic with much of what Rupert says, particularly his attempts to undermine the arguments for HEC, his positive view of systems boundaries seems vulnerable to several objections.

The details of his approach can be stated with some formal precision (Rupert, 2009, pp. 42-3). Take all of the token performances of cognitive tasks that a subject has executed—e.g., visually identifying an object, constructing a plan, remembering items from a list. Each of these task performances (t_1, t_2, …, t_n) involves a whole host of cognitive processes of various sorts. For each performance, construct the set consisting of the processes that were involved in it; e.g., t_3 involved only processes {A, B, C}. For each process involved in a set, determine the conditional probability of its co-employment with the other processes that are included in that set: P(A|B&C), P(B|A&C), P(C|A&B), where these values need not be equal. Now construct a single list of all of the sets of co-activated cognitive processes across all tasks, rank-ordered by these probabilities. Because a single set can be assigned several non-equal probabilities, the same set can occur in many places on this list: {A,B,C} may occur at 0.7, 0.5, 0.3, etc. The result of this procedure is a list of all cognitive processes used in all cognitive tasks that a subject has performed, ordered by the degree to which they are co-involved in performing these tasks.

Now divide this list in two, either at the 0.5 mark or at any other seemingly natural-looking gap. Carve off the top section of the list; this constitutes a new list of each set of processes or mechanisms that is highly co-involved. Count the number of times each highly co-involved process has occurred on this list. So A might score 25, B might score 23, C might score 3, etc., depending on how many sets they occur in. Finally, locate another natural cutoff point separating highly reused from infrequently reused components. The high-scoring components on this list are part of a single system, while the low-scoring ones may be resources used by the system, but not part of it.

The governing idea of this procedure is to identify systems with sets of processes that are (1) highly co-involved and (2) frequently re-used in the production of cognitive outcomes. Rupert concedes that this may not *constitute* something's being a cognitive system, but 'so long as a subject has a fair amount of experience in the world' (p. 43), it is at least diagnostic of systems integration. There is undeniably something right in the idea that systems are highly co-involved sets of processes—indeed, the description of cognitive systems I am operating with presents them as sets of interlocking mechanisms for producing such processes. But Rupert's conditions face several problems.

One cluster of objections centers around the various potentially non-objective decisions that need to be made in employing Rupert's criteria. First, it is notoriously hard to decide what counts as a 'task' in cognitive science, and constructing these lists depends on some relatively determinate notion of when one task has been performed. But since I have no good answer to the problem of task individuation (and doubt whether anyone else does), set this aside. Second, we need to decide what is an acceptable co-occurrence value to partition the first list, and what is an appropriate frequency of re-use to partition the second. Here too it is hard to settle on objective

criteria for making these decisions. But given that there are also somewhat arbitrary choices made elsewhere about what counts as good evidence in science (e.g., the choice of a standard significance value of $p<0.5$), I won't emphasize this either.

The more basic problem with Rupert's criteria is that we should allow that there can be parts of a cognitive system that are in fact relatively isolated and rarely used. Two examples will make the point. Complex computer programs are typically decomposed into a host of subroutines. Good programming practice mandates that frequently used operations should be separated out into their own subroutines, both to reduce redundancy in the code and to enable these routines to be separately debugged and optimized. But even a subroutine that is rarely or never called in the history of the program's being run is still part of the program, and hence part of the system that is running the program.[11] For an extreme example, consider a 'suicide subroutine' that is designed to wipe the system's storage and memory in case of emergency. This routine functions in one and only one task that might never be run—or if it is run, it is certainly only run once—but it is a part of the system that contains it nevertheless.

For another example, notice that what is part of a cognitive system on this account depends on the range of experience the subject has in the world; that is, on the history of the tasks she has performed. There are both real-life and fictional cases of individuals with extremely impoverished experience, however, whose cognitive systems possess untapped capacities. The case of Genie, who was raised in a profoundly linguistically impoverished environment, is one example (Rymer, 1994). Assuming the worst about her circumstances, she might have activated processes pertaining to language acquisition relatively few times. On Rupert's index, these might not count as part of her cognitive system. Or consider a human

[11] Sometimes programmers code these routines as jokes, known as 'Easter eggs', to be triggered only when an arcane set of commands is entered.

raised without the benefit of formal mathematical education. He might never use those cognitive capacities that enable us to perform arithmetic, but his cognitive system presumably contains processes that would allow him to do so (Dehaene, 1999). The moral here is that even subjects without a fair amount of experience in the world can possess cognitive systems having untapped capacities.

These problems arise from the fact that Rupert's criterion constructs cognitive systems from the cognitive processes used in the real-world history of a subject. Rupert correctly notes that, intuitively, what is problematic about HEC is that it involves what Clark (2008, p. 158) calls 'transient extended cognitive systems': temporary organism-world ensembles that come together to solve local problems and then dissolve. Appeal to real-world history of use will rule these out as being part of my cognitive systems, since they score low on the index, but it will rule out much else as well. A natural move to make is to appeal to the processes that subject *would* be disposed to use in various other circumstances, understood as either alternative past histories or synchronically defined counterfactual scenarios. This solves the Genie case and the case of the mathematical illiterate; in the right circumstances they *would* have used the processes and capacities that we are inclined to ascribe to them. However these possibilities are defined, however, they will need to avoid simply reintroducing the problem of causal spread and cognitive bloat that plague HEC.

To see this, note that if I were placed in the right circumstances, I might have no choice but to rely on external props and aids—if, for instance, I were chained to a desk and forced to perform menial arithmetic calculations on paper for the rest of my life. Given that this external process is something that I *would* make extensive use of in certain circumstances, a dispositional criterion would have to include it as part of my cognitive system. Similar cases can easily be

generated, since we can always come up with circumstances in which any subject would rely on external resources much more frequently and for a much greater amount of time than we actually do. Without restrictions on the range of circumstances that we are allowed to take into account in defining these dispositions, Rupert's criterion will not do the work of defining systems in a way that excludes HEC on principled grounds.

Where the approach goes wrong, I suggest, is in taking the boundaries of systems to be defined by the frequent re-use of the same body of processes. Cognitive systems are composed of interlocking sets of mechanisms that produce such processes, but how often they are actually used, and how frequently they co-occur in various task performances depends in unpredictable ways on the history and environment of the cognizer. If anything, the order of definition should go in the other direction. We should start by specifying the mechanisms that are embodied in a particular system, then spell out the processes that occur within that system, rather than starting with processes and hoping to build up systems from them. Rupert's index of integration is, then, at best a guide to system boundaries in a fuzzily-specified range of normal conditions. The transducer view, by contrast, provides a criterion that applies across a wide range of environments and histories, both normal and abnormal.

6. Terminators, Martians, and the Parity Principle

A possible objection to the transducer view is that relying on transducers to delimit cognitive systems is question-begging in the context of debates over HEC. Clark & Chalmers (1998) in effect consider a version of this objection based on perception, rather than transduction: 'From the standpoint of [the Otto-notebook] system, the flow of information from notebook to brain is not perceptual at all; it does not involve the impact of something outside the

system. It is more akin to information flow within the brain' (p. 16). Transducer boundaries are irrelevant because the process taking place between Otto's biological brain and his notebook is one that could just as well take place within a creature's head. The fact that transduction (or perception) is involved is irrelevant.

To sharpen and motivate the claim that even intracranial information retrieval may involve 'perceptual phenomenology', and hence that this phenomenology doesn't disqualify extracranial retrieval from being part of a creature's cognitive processing, they offer the Terminator counterexample. When the Terminator retrieves information from memory, it appears as text in his visual field. Presumably this text is read by him and used to guide his murderous actions. But this is a purely internal process, and it is plausibly cognitive. The Parity Principle claims that spatial location is irrelevant to deciding whether something is a cognitive process or not; all that matters is whether the process realizes the right functional structure. So the externalized analog of this process should count as a cognitive process as well.

The Terminator doesn't constitute a counterexample to the transducer view, however. Indulging in some speculation, I'd imagine that the Terminator's text display is produced endogenously by some sort of visual imagery process. Otto's reading, by contrast, is transducer mediated. The fact that the visual image produced in each case is identical shouldn't lead us to group these cases together. One might protest that in both cases there is an information store (external for Otto, internal for the Terminator), a textual representation that is produced by that store (the visual image in each case) and a process that interprets this visual image it to extract its content (again, the same process in each case). The difference is that in the Otto case, as opposed to the Terminator case, the process of moving from the information store to the representation is transducer-mediated; the textual image is produced by the transduction of physical signals into

representations. No such process takes place in the Terminator case. So we can save the intuition that the Terminator has a merely eccentric way of accessing its standing beliefs and memories, but deny that Otto's book contains his.

Sprevak (2009), in the same vein, argues that if we accept the possibility of various kinds of Martians, we are also committed to HEC. If we are wondering whether to count information stored in the form of ink marks on paper as part of my cognitive system, we should imagine a Martian who stores such marks on an intracranial scroll and then transforms them into a bit-mapped representation which feeds into further cognitive processes. We would not want to say that the Martian is not cognizing merely because her thinking happens in an eccentric way—that would be unacceptably chauvinistic. So by the Parity Principle, again, we should count the environmentally extended system as cognitive.

Whether we accept this inference depends on how the Martian case is described. In particular, it depends on whether the internal ink marks are playing a representational role in her thinking. What it means to play a representational role in this case is that the marks comprise a subset of the domain of some operation that takes representations and transforms them into other representations. Playing a representational role is being subject to a set of processes and mechanisms in virtue of *being a representation* of a certain type.[12] This is a matter of being consumed or produced by a mechanism whose function is described in terms of transformation on content-bearing structures. If there is a process that is causally driven by the presence of a certain type of representation, including these ink marks, then those marks play such a role.

[12] Compare Ramsey's description of how what he calls IO (Input-Output)-representations function (Ramsey, 2007, pp. 76-7): 'computational processes treat input and output symbolic structures a certain way, and that treatment amounts to a kind of job assignment—the job of standing for something else. […] Serving as a representation in this sense is thus to function as a state or structure that is used by an inner module as a content-bearing symbol.' Transducer inputs are not used as content-bearing symbols, or as symbols at all, even if they *happen* to be symbols. The representations manipulated in mental arithmetic, however, are used in this way by the processes that operate over them.

If, on the other hand, the marks are not playing their role in virtue of their representational properties but rather in virtue of, say, their formal or physical properties, then they do not count as playing a representational role. The marks might comprise a subset of the domain of some mechanism that takes physically characterized inputs and produces representations as output; that is, they might be inputs to a transducer. But while transducers may operate on representations sometimes, they don't take them as their input in virtue of their having any particular representational properties. A set of marks on paper is input to a visual transducer whether it is meaningful or meaningless. It doesn't play the role it does in virtue of being a certain kind of representation. This fact is clear once we note that the representational properties of the marks, if they have any, are 'invisible' to the transduction process; it produces a representation of the physical properties of the marks, but *their* representational content is not part of what is produced or manipulated.

Now consider two ways of building one of Sprevak's Martians. In one way—the way he seems to describe—the ink marks themselves are transduced by a process that produces a bitmap image, which is then manipulated in various ways. In this case, they are not playing a representational role in the present sense. They are mere inputs to the cognitive system, albeit inputs that happen to be internally stored. If, on the other hand, the ink marks themselves are manipulated by processes that operate over them in virtue of their being representations of the right type, then they are playing the right sort of role, and should count, all else being equal, as part of the cognitive system.[13]

[13] To return to a distinction made in section 4, processes that cross interfaces are ones in which one mechanism produces representations that serve as representations *for* another process. The CPU produces computational instructions for the graphics card that are transmitted via the bus; the graphics card does not have to transduce these signals by representing their electrical properties. Those electrical signals just *are* serving as representations for the circuitry onboard the card. The difference in the present case involves whether the Martian's ink marks are more like *physical media to be represented* or *physical media functioning as representations*.

Advocates of HEC are likely to object by pointing out these differences are simply irrelevant. The fact is that we have a representation that is being manipulated as part of the process of coming to solve a cognitive task, i.e., performing long division, retrieving some fact from memory, and so on. It interacts with other representations and processes in so doing. That this representation is extrabodily should pose no objection in principle. Why should the mere fact that this causal interaction crosses transducer-effector boundaries somehow render this external vehicle not part of a single, ongoing cognitive process that spans body and world? Doesn't the Parity Principle, and the more basic functionalist thesis that underlies it, counsel us to ignore these facts as irrelevant matters of implementation?

I think that this objection rests on an unduly expansive notion of what functionalism entails. Functionalism simply claims that psychological states, processes, and systems are to be functionally individuated (rather than physically, chemically, etc.). I assume that both advocates and opponents of HEC are functionalists in this sense.[14] But the issue is *which* functionalist specification of cognitive processes and systems we should adopt. The transducer-effector view is thoroughly functionalist: being a transducer or effector is something functional, not physical. These are location-independent, realization-independent properties, and that is all that functionalism or the Parity Principle requires. No questions are being begged against HEC by this view.

If the question is why we should prefer this functionalist specification of systems to the ones proposed by HEC, the response is that the HEC specifications, insofar as they are explicitly spelled out, result in bloating cognitive systems unacceptably. I have given several examples of this in section 4; the case has also been laid out in detail by Sprevak (2009). He argues that the kind of functionalism embodied in the Parity Principle (which he calls the 'fair-treatment

[14] See Shapiro (2008) for more discussion of why functionalism *per se* is neutral with respect to HEC.

principle') is sufficiently unconstrained that it leads to such unacceptable results as my suddenly believing all of the contents of the books in a library upon my entering it, and my having a cognitive capacity for calculating dates in the Mayan calendar once I install such a program on my laptop.[15] Because all of these external resources are available to us, and because we could interact with them in a way that involves manipulating information to achieve some cognitive goal or carry out some cognitive task, we are committed to saying that they are part of our cognitive systems, or at least that we are a part of a vast cognitive system that also includes them as parts.

The undesirability of this conclusion is, I take it, obvious. Let me make one further point in defense of the approach taken here. Sprevak holds that we should be committed to a form of functionalism sufficiently liberal to permit Martians with bizarre realization to be genuine cognitive agents. Agreed: not all cognition is internally organized in the way that human cognition is. The transducer view permits this insofar as it places effectively no constraints on the sorts of cognitive processes a creature might possess. It sets fairly clear boundary conditions on these creatures, however. So the functional specification for being a cognizer may allow for wide latitude in realization, but radical forms of HEC do not necessarily follow from this. The library, for example, does not become part of my cognitive system upon my entering it. Neither is there any *larger* system that comprises both me and the library, or me and my laptop. The reason is that there is no set of transducers and effectors such that all of these representations and resources are contained within them, in the sense spelled out in section 2.

7. Conclusions

[15] The example of acquiring a whole set of beliefs on setting up camp in a library was also advanced by Rupert (2004), who gives many further examples of cognitive bloat along these lines. See also Rupert (2009), pp. 15-35.

I suggest that the transducer-effector view is just the sort of thing Goldilocks is looking for: an explicit, principled, functionalist-friendly criterion that underlies much of the successful work in cognitive science to date. Notice as well that there is nothing in the view that rules out extended cognition in principle. Extended computational systems, as noted earlier, are already a reality. If brains could be connected to external media by mechanisms that contain operations mapping neural representations onto, say, silicon representations via interfaces, we would have actual examples of this. We already have, amazingly enough, examples of neural prostheses that constitute artificial transducer and effector systems: robot arms that can be controlled by the thoughts of monkeys, and silicon retinas that can interface with visual cortex. The architecture that our brains implement, then, appears to be interface-ready.

However, as exciting as these developments are, we should keep our heads. There are structures that look for all the world like *natural* boundaries between cognitive systems and the environment. These are the places where the (merely) physical becomes representational, and where representations in turn become (merely) physical. These transition points constitute the bounds of cognitive systems. And these systems can be intricately embedded in webs of supportive causal and informational interaction with their environment without thereby incorporating that environment.

Acknowledgments

Thanks to Ken Aizawa and an anonymous referee for helpful comments on an earlier version of this paper.

References

Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. Malden, MA: Blackwell.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Perception and Cognition* (pp. 261-325). Minneapolis: University of Minnesota Press.

Clark, A. (2008). *Supersizing the Mind*. Oxford: Oxford University Press.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58, 10-23.

Dehaene, S. (1999). *The Number Sense*. Oxford: Oxford University Press.

Dennett, D. (1978). Where am I? In D. Dennett, *Brainstorms* (pp. 310-323). Cambridge: MIT Press.

Gershon, M. (1998). *The Second Brain*. New York: HarperCollins.

Haugeland, J. (1998). Mind embodied and embedded. In J. Haugeland, *Having Thought* (pp. 207-237). Chicago: University of Chicago Press.

Pylyshyn, Z. (1984). *Computation and Cognition*. Cambridge: MIT Press.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Robbins, P., & Aydede, M. (2008). *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press.

Rowlands, M. (1999). *The Body in Mind*. Cambridge: Cambridge University Press.

Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology*, 22, 1-19.

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

Rupert, R. (2009). *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Rymer, R. (1994). *Genie*. New York: HarperCollins.

Schneider, S. (2009). The nature of symbols in the language of thought. *Mind and Language,* 24, 235-555.

Shapiro, L. (2008). Functionalism and mental boundaries. *Cognitive Systems Research*, 9, 5-14.

Sprevak, M. (2009). Extended cognition and functionalism. *Journal of Philosophy*, 106, 503-527.