THE JOURNAL OF PHILOSOPHY
VOLUME CI, NO. 8, AUGUST 2004

CHALLENGES TO THE HYPOTHESIS OF EXTENDED COGNITION*

In recent decades, an intriguing view of human cognition has garnered increasing support. According to this view, which I will call the *hypothesis of extended cognition* (HEC, hereafter), human cognitive processing literally extends into the environment surrounding the organism, and human cognitive states literally comprise—as wholes do their proper parts—elements in that environment; in consequence, while the skin and scalp may encase the human organism, they do not delimit the thinking subject.[1] The hypothesis of extended cognition should provoke our critical interest. Acceptance of HEC would alter our approach to research and theorizing in cognitive science and, it

* I would like to thank Aaron Meskin, David Chalmers, Philip Robbins, and Josh Osbun for comments on earlier drafts of this essay. Shorter versions of the paper were presented at the Mountain-Plains Philosophy Conference (2001), Texas Tech University, and the 2002 meeting of the Pacific Division of the American Philosophical Association. I would like to express my appreciation to members of all three audiences for their useful feedback (especially William Lycan at the Mountain-Plains and David Chalmers at the APA), as well as to my conference commentators, Robert Welshon and Tadeusz Zawidzki. I also benefited from discussing extended cognition with Edward Averill, Dan Kaufman, Nick Rupert, and Jay Newhard.

[1] See Andy Clark and David Chalmers, "The Extended Mind," *Analysis*, LVIII (1998): 7–19; Mark Rowlands, *The Body in Mind: Understanding Cognitive Processes* (New York: Cambridge, 1999); Clark, *Being There: Putting Brain, Body, and World Together Again* (Cambridge: MIT, 1997), "Reasons, Robots, and the Extended Mind," *Mind and Language* XVI (2001): 121–45; John Haugeland, "Mind Embodied and Embedded," in Y. Houng and J. Ho, eds., *Mind and Cognition* (Taipei, Taiwan: Institute of European and American Studies, Academia Sinica, 1995), pp. 3–37; Tim van Gelder, "What Might Cognition Be, If Not Computation?" this JOURNAL, XCII, 7 (July 1995): 345–81; Daniel C. Dennett, *Kinds of Minds: Toward an Understanding of Consciousness* (New York: Basic Books, 1996); Ruth Garrett Millikan, *White Queen Psychology and Other Essays for Alice* (Cambridge: MIT, 1993), essays 7–9. Work outside of academic philosophy has provided much of the conceptual impetus for HEC: Edwin Hutchins, *Cognition in the Wild* (Cambridge: MIT, 1995); Richard Dawkins, *The Extended Phenotype* (New York: Oxford, 1982); Merlin Donald, *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition* (Cambridge: Harvard, 1991).

0022-362X/08/0103/000–000

would seem, significantly change our conception of persons. Thus, if HEC faces substantive difficulties, these should be brought to light; this paper is meant to do just that, exposing some of the problems HEC must overcome if it is to stand among leading views of the nature of human cognition.

The essay unfolds as follows: the first section consists of preliminary remarks, mostly about the scope and content of HEC as I will construe it. Sections II and III clarify HEC by situating it with respect to related theses one finds in the literature—the hypothesis of embedded cognition and content externalism. The remaining sections develop a series of objections to HEC and the arguments that have been offered in its support. The first objection appeals to common sense: HEC implies highly counterintuitive attributions of belief. Of course, HEC theorists can take, and have taken, a naturalistic stand. They claim that HEC need not be responsive to common-sense objections, for HEC is being offered as a theoretical postulate of cognitive science; whether we should accept HEC depends, they say, on the value of the empirical work premised upon it. Thus, I consider a series of arguments meant to show that HEC is a promising causal-explanatory hypothesis, concluding that these arguments fail and that, ultimately, HEC appears to be of marginal interest as part of a philosophical foundation for cognitive science. If the cases canvassed here are any indication, adopting HEC results in a significant loss of explanatory power or, at the very best, yields only an unmotivated reinterpretation of results that can, at little cost, be systematically accounted for within a more conservative framework.

I. PRELIMINARIES

First, let us hear from HEC's proponents. Mark Rowlands offers the following as one of the two primary theses of his recent book, *The Body in Mind*: "Cognitive processes are not located exclusively inside the skin of cognizing organisms" (*op. cit.*, p. 22). And from this, Rowlands infers, "[I]f we assume that the mind of a cognizing organism such as a human being is made up, at least in part, of cognitive processes, the central metaphysical assertion of this book is that the mind is not, exclusively, inside the head" (*op. cit.*, p. 29).[2] Andy Clark and David Chalmers also give HEC fairly clear expression: "In particular, we will argue that *beliefs* can be constituted partly by features of the environment, when those features play the right sort of role in driving

[2] And such spatial-sounding talk appears to be meant in precisely that way—see pp. 44–45.

cognitive processes. If so, the mind extends into the world" (*op. cit.*, p. 12). Or as Clark has more recently asserted, "The intelligent process just *is* the spatially and temporally extended one which zig-zags between brain, body, and world."[3]

I initially characterized HEC in terms of *cognitive* states and processes, rather than in terms of *mental* states and processes, but also at issue is the extension of the *mind* into the environment beyond the individual human organism. In the discussion that follows, we will keep one eye on this bigger issue. Cogent criticisms of HEC will not, of course, refute the hypothesis of an extended mind; given, however, that current work on extended cognition promises to provide the strongest support to date for the view that the mind is extended, HEC's problems are, in no small measure, problems for proponents of extended minds.[4]

Herein I evaluate HEC only as it applies to individual subjects traditionally conceived of, that is, to subjects that are not individually composed of more than one mind. This excludes consideration of what Edwin Hutchins calls "socially distributed cognition" (*op. cit.*, p. 129), Hutchins's central example of which is a team's navigation of a large ship.[5] Hutchins argues that, in such cases, we rightly attribute mental states to groups as single units; such mental states include remembering, perceiving, having expertise, and entertaining hypotheses and being biased in the evaluation of them.[6] What is more, Hutchins claims that in many of these cases, the group-system, as a whole, instantiates a cognitive or mental state that no individual member of the group instantiates.

We should not dismiss the possibility that a group of organisms, each of which is the locus of a mind, can be the subject of a cognitive

[3] "Reasons, Robots, and the Extended Mind," p. 132.

[4] Given that at least some aspects of one's mind constitute central parts of one's self, the present debate would seem to bear also on our understanding of the self. Clark and Chalmers close their article with a discussion of just this issue: "What, finally, of the self? Does the extended mind imply an extended self? It seems so.... Once the hegemony of skin and skull is usurped, we may be able to see ourselves more truly as creatures of the world"—p. 18; also see Clark, *Being There*, pp. 213–18 (although note that, partly because of a concern about agency and moral responsibility, Clark sometimes resists the extended view of the self—"Time and Mind," this JOURNAL, XCV, 7 (July 1998): 354–76, see p. 367). Impugning HEC does not, of course, disprove the extended view of the self; but as in the case of the extended mind, criticisms of HEC strike a blow to the view, for they speak against what is offered as one of the strongest reasons to embrace the view that the self is extended.

[5] See especially chapters 4 and 5.

[6] Hutchins, p. 196, regarding memory; pp. 182, 194, regarding perception/detection; chapters 4 and 5, passim, regarding expertise at navigation; pp. 239–61, regarding hypothesis-testing, interpretation, and confirmation-bias.

or mental state. Nevertheless, if we are to infer an extended mind or self from the existence of extended cognitive states, we should begin by evaluating arguments for HEC that take, as their starting point, the clearest cases of mental-cum-cognitive activity—cognition as it appears in something close to the individual subject. If the best way to explain the individual self's intelligence requires that proper parts of her mental-cum-cognitive states extend into the extraorganismic environment, then the case for an extended mind looks fairly strong. If, instead, our inference to extended minds begins with observations about group intelligence, our grip on the issues seems less secure. Perhaps Hutchins is right: groups act as computational systems, of a sort. Whether this form of computation underlies cognitive processes that best explain the *mental* states or capacities of groups is another matter; it depends, to a great extent on whether groups *have* mental states or capacities, and here many of us, quite rightly, lose our bearings. Thus, given our ultimate interests, in extended minds and selves, it will be most fruitful to begin, as HEC-inclined philosophers of mind and cognitive science typically have, with an examination of paradigm cases of mental capacities and activities—those of individual minds.

We should also be clear about HEC's strength. Sometimes HEC theorists make only tentative commitment to HEC as a claim about actual human cognition; instead, they offer a fairly weak modal claim, to the effect that there is a possible world in which some cognition is extended. Rowlands states as his primary goal merely to loosen the grip of the internalist view, to make it easier for us even to conceive of the cognitive or mental in extended terms (*op. cit.*, pp. 12, 15, 149, 172).[7] Given, however, the empirical nature of many of the HEC theorists' arguments and the extent to which HEC theorists take themselves to be contributing to the foundations of cognitive science, I treat HEC as a substantive hypothesis about a significant amount of human cognition. I take for granted the weaker modal claim, that extended cognition is possible, without meaning to suggest that this claim is uninteresting or unworthy of further discussion. This approach is warranted by the pursuit of genuinely important intellectual goals: to figure out whether HEC, in its more substantive form, is true, or at least whether it provides a promising framework within which to study cognition and mind as they appear in the actual world. If HEC

[7] Clark and Chalmers sometimes offer what appears to be a similar take: "The moral is that when it comes to belief, there is nothing sacred about skull and skin" (p. 14); furthermore, in their responses to various objections, Clark and Chalmers often point out the in-principle possibility that external resources function as parts of a cognitive system.

does not provide a promising framework for the pursuit of cognitive science (as it attempts to understand actual mental states), the radical theses of extended mind and extended self lose much of their current appeal; one cannot infer that the human mind or self is extended—or that we are creatures of the extended world—from a premise asserting the bare possibility of extended cognition.

II. EMBEDDED COGNITION AND HEC

The cognitive activity of a subject, individuated by organismic boundaries, consists at least partly in the thinker's interaction with her environment. This may seem an uncontroversial point, but its degree of triviality is inversely proportional to the *degree* to which the thinker is claimed to exploit the environment in her cognitive activity. According to the *hypothesis of embedded cognition* (call it HEMC), cognitive processes depend *very* heavily, in hitherto unexpected ways, on organismically external props and devices and on the structure of the external environment in which cognition takes place.[8] Adopting such a view significantly affects our estimation of what goes on inside the thinking subject—for example, which computations she must perform using her own neural resources in order to exercise a given cognitive ability. Of further moment in the present context is the natural way in which one might infer HEC from HEMC: recognizing the extent to which cognition depends on, for example, the manipulation of environmental props tempts one to include those external props as mereological parts of the subject's cognitive states or processes; the external elements participate intimately in cognitive processing, so why not think that changes in their states constitute *part of* that processing?[9] The degree of cognitive co-activity among the thinking or-

[8] See Ron McClamrock, *Existential Cognition: Computational Minds in the World* (Chicago: University Press, 1995).

[9] For a suggestion along these lines, see Clark, *Being There*, p. 105ff. Clark proposes a style of explanation he calls "catch-and-toss," which, like HEMC, preserves a clear boundary between the subject as organism and the environment; however, Clark goes on to argue that the actual relationship between subject and environment is more complex and interactive than is allowed by the catch-and-toss model, so much so that we should embrace HEC.

Dennett appears to make a similar move (pp. 135ff.). To illustrate the way in which external marks can serve as cognitive aids, Dennett explains why some elderly people can function at home but not in an institution: at home they have numerous "landmarks," "triggers," and "reminders" (p. 138) that keep their habits of self-care on track—a sensible enough HEMC-style explanation. Dennett then recasts his point in terms of an extended mind (or infers an extended mind from the bruited HEMC-style considerations—the exposition is unclear): he claims that when we remove such elderly people from their homes, we are "literally separating them from large parts of their minds" (p. 139).

This ambiguity of message also appears in work in cognitive science, for example, in J. Kevin O'Regan's discussion of visual perception; see "Solving the 'Real' Mysteries

ganism and elements in her environment creates an interactive system that is, one might conclude, most fruitfully viewed as a single unified system. As Hutchins puts it, "When we turn to the coordination events and see all the media that are simultaneously in coordination (some inside the actor, some outside), we get a different sense of the units in the system" (*op. cit.*, p. 158), and "Again, the normally assumed boundaries of the individual are not the boundaries of the unit described by steep gradients in the density of interaction among [representational] media" (*op. cit.*, p. 157). We will return presently to the question of whether a robust version of HEMC implies HEC; let us first get a better grip on HEMC.

Consider a simple example of the sort of cognitive strategy typically taken to manifest embedded cognition. A subject is in a room with only one other macroscopic, animate object: his best friend. Verbal communication with the friend might proceed in the following way: the subject wishes to speak to the friend. He scans the room and, prior to a change in orientation and the onset of speech, he verifies that a certain "object" in the room is in fact the best friend; to do so, he refers to a continuously maintained detailed internal image (or description) of the various objects in the room and searches that scene (or list) for an object matching the detailed pictorial (or propositional) representation of his best friend, called up from long-term memory. For obvious reasons, this manner of reidentification makes heavy computational demands on the cognitive system. In contrast, an embedded account identifies ways in which the visual system simplifies the internal computational problem by, among other things, allowing contingent facts about the specific environment to guide cognition. On this view, the subject need not maintain a detailed internal model of the room in order to orient himself toward the intended addressee. Instead, he gets by on the information that his best friend is the only other easily visible, animate object in the room; a quick scan for signs of life, sans either the construction or maintenance of a detailed model, allows timely orientation toward her. (Certain further complications must be taken into account, for example, the subject must in some way represent that activity near the door requires further

---

of Visual Perception: The World as an Outside Memory," *Canadian Journal of Psychology*, XLVI (1992): 461–88. O'Regan sometimes describes visual processes in a way that suggests embedded cognition: he claims that the visual system quickly collects information from the environment when the subject needs it, rather than maintaining a scale-model of the surrounding environment (pp. 470–71), apparently leaving intact the privileged status of the organism as cognitive processor. As O'Regan's title suggests, though, he sometimes advocates a version of HEC, by proposing that the environment functions as *part of* the subject's memory (pp. 472–73).

inspection. But even having added the necessary supplementary beliefs—or belief-like states—the operative simplifying assumption greatly reduces the subject's internal computational load.)[10]

Notice the extent to which HEMC's acceptance alters the cognitive scientist's research strategy: in the case of visual reidentification, investigation shifts toward the study of ways in which the visual system efficiently exploits relevant environmental cues, and away from the attempt to figure out how the subject constructs a detailed, continuously updated, internal model of the surrounding environment. Typically, then, a cognitive theory premised upon HEMC will posit less representational and computational structure internal to the subject.[11] Nevertheless, HEMC is significantly less radical than HEC. According to HEMC, we can properly understand the traditional subject's cognitive processes only by taking into account how the agent exploits the surrounding environment to carry out her cognitive work. In contrast, HEC implies that, for many purposes, we should set aside our focus on the traditional subject: the unit of analysis should be the organism and certain aspects of its environment treated together, as a single, unified system.

In what follows, then, I treat HEMC and HEC as offering distinct, competing explanations of various cognitive phenomena. Of great dialectical importance will be the question whether we can make do with HEMC, or whether HEC offers superior explanations of the phenomena of interest to cognitive scientists. If HEC does not, then all other things being equal, we should endorse HEMC over HEC, by dint of the methodological principle of conservatism.

Why, though, might it have seemed that HEC follows from HEMC? One common, yet objectionable, argument for HEC rests on a claim of epistemological dependence: we cannot fully understand human

[10] This example illustrates the idea of task-directed perception; see McClamrock, pp. 134–38; Clark, "Moving Minds: Situating Content in the Service of Real-time Success," in James E. Tomberlin, ed., *Philosophical Perspectives, 9: AI, Connectionism, and Philosophical Psychology* (Atascadero, CA: Ridgeview, 1995), pp. 89–104, especially pp. 97–98. Also see Patricia S. Churchland, V.S. Ramachandran, and Terrence J. Sejnowski, "A Critique of Pure Vision," in Christof Koch and Joel L. Davis, eds., *Large-Scale Neuronal Theories of the Brain* (Cambridge: MIT, 1994), pp. 23–60; these authors argue, as O'Regan does, that at any given time the subject views only a "visual semiworld" (p. 25), rather than a full—or anything close to a full—model of the surrounding environment.

[11] Rodney Brooks takes HEMC-style (among other) considerations to imply a radically deflationary view of what goes on inside the subject; see, "Intelligence without Representation," in Haugeland, ed., *Mind Design II: Philosophy, Psychology, Artificial Intelligence* (Cambridge: MIT, 1997), pp. 395–420.

cognition unless we consider the context in which it is embedded,[12] and thus the embedding context must be part of cognition itself. The epistemic advice on offer seems sensible enough and well supported by many HEMC-style examples, but it hardly shows that there are cognitive states of systems that individually include a single human organism and some of the elements of her environment. Compare: one wishes to understand an important historical event, say, Nazi Germany's invasion of Poland. In order fully to understand this event as an historical event, one would need to know, among many other things, a great deal about the economic conditions in Germany during the nineteen-twenties. This does not imply that the economic conditions in Germany during the nineteen-twenties are *part of* the invasion. It is simply false that whenever a full understanding of *A* includes some cognizance of its relations to *B*, *B* is a mereological part of *A*. The argument from epistemological dependence given above lacks a premise, and it is difficult to see what plausible candidate one might add. Perhaps the HEC theorist has something like the following principle in mind: in any case where cognizance of *A*'s relation to *B* is significantly relevant to our understanding of *A*, we should posit a system, *A-B*, as a single unit of study. As stated, the principle lacks clear content; a number of aspects of its application must be clarified, including how to treat overlapping systems, how to individuate systems over time, and *how relevant A*'s relations to *B* must be to our understanding of *A* in order to justify positing a single *A-B* system. Depending on how we resolve these questions, such a principle might saddle us with an unacceptable proliferation of systems (many of them extremely short-lived) or narrow study to a very small number of cognitive systems, perhaps only one (because, after all, everything is related to everything else, at least in some indirect way). Although we should not dismiss these possibilities out of hand, the utility to the study of cognition of such system-taxonomies must be established by something other than mere epistemological relevance.

Worthy of more attention is an argument, of sorts, outlined earlier: it is not that HEMC directly supports HEC. Rather, as one examines more and more closely the complex, cognition-sustaining interactions between organism and environment, it becomes apparent that the very distinction between organism and world is unmotivated. Such examination eventuates in a flash of paradigm-shifting insight that

[12] This claim is made by theorists explicitly motivated by HEMC—for example, Hutchins, pp. 169, 290—as well as by some theorists inclined toward HEC but motivated by concerns other than the embedded nature of cognition—for example, Millikan, p. 181.

reveals the empirical power to be gained by embracing HEC and leaving behind HEMC's commitment to an important theoretical distinction between organism and environment. The discussion below focuses at some length on the question of whether such a reconception yields empirical pay dirt[13]; for now, however, we can recognize the way HEMC-style thinking might inspire HEC theorists, while insisting that HEC stand as a competitor to HEMC that must be evaluated on its own merits.

III. CONTENT EXTERNALISM AND HEC

Content externalism and HEC are distinct, though mutually consistent, theses: neither HEC nor its negation follows from content externalism, and HEC does not entail content externalism.[14] There is, however, genuine risk of misunderstanding regarding this matter: externalist theories of content, mental as well as linguistic, have had enormous influence on recent philosophical thinking in the Anglo-American world[15]—so much so that some readers might assume that HEC is just another way of stating the externalist view (or perhaps one of its important implications). In the interest of clarity, then, this section examines the relation between the two views, explaining why I treat HEC independently of the sort of issues normally addressed in discussions of content externalism.

As applied to the mental, content externalism holds that the content of at least some mental states is determined at least partly by the

[13] Dawkins introduces his discussion of the extended phenotype with a caveat: he is mostly out to "change the way we see" data and facts (p. 2), where 'to see' means to interpret. In this spirit, one might insist that I am skewing the discussion by evaluating HEC in terms of an empirical payoff—as if we could gauge such success at present. Note, however, that Dawkins's strategy cuts both ways: HEC theorists adduce empirical considerations as sight-shifting inspiration, admitting that these considerations fall far short of an empirical demonstration that HEC will carry the day in cognitive science. Similarly, although the empirical considerations of the present essay may not prove that HEC will lose out to a more traditional approach, they would seem, at the very least, to provoke reasonable resistance to the HEC theorist's proposed paradigm shift.

[14] In what follows, I address exclusively the question whether content externalism implies HEC, arguing that it does not; but it is important to see why the converse relation also fails: we might decide it best to conceive of cognitive systems as extended, while endorsing a nonexternalist theory—for example, a conceptual-role account—of the content of those systems' extended states.

[15] Tyler Burge, "Individualism and the Mental," in Peter A. French, T.E. Uehling, and Howard K. Wettstein, eds., *Midwest Studies in Philosophy, Volume 4: Studies in Metaphysics* (Minneapolis: Minnesota UP, 1979), pp. 73–121, "Individualism and Psychology," *Philosophical Review*, XCIII (1986): 3–45; Hilary Putnam, "The Meaning of 'Meaning'," in *Mind, Language, and Reality: Philosophical Papers*, Volume 2 (New York: Cambridge, 1975), pp. 215–71; Saul Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980).

subject's relations to kinds or individuals in her external environment. The externalist view is often stated thusly: two subjects could be molecule-for-molecule duplicates of each other, yet be in mental states with different contents—this in virtue of differences in the subjects' relations to their physical or social (that is, extraorganismic) surroundings. If these ways of putting matters seem obtuse, one might recall Hilary Putnam's more pithy and colorful formulation of linguistic content externalism's main lesson: "'[M]eanings' just ain't in the *head*!" (*op. cit.*, p. 227). The distance between content externalism and HEC may now seem clear: it is one thing to say, as content-externalists often do, that the contents of mental states, and thus the mental states themselves, are *individuated* partly by the relations those states bear to certain individuals, kinds, or practices in the subject's environment; it is quite another to say, as HEC theorists do, that elements in the organism's environment appear as *mereological constituents* of the thinking subject, her cognitive states, or cognitive processes.

Other ways of motivating content externalism might, however, reveal a tighter link between HEC and the externalist view. A more Russellian approach to thought content, for example, would seem not only to imply a form of content externalism but also to make concrete individuals parts of the contents of some belief states, thus smearing out subjects' minds into the surrounding physical environment. This is most clearly the case for *de re* beliefs,[16] which are often expressed by demonstrative constructions, for instance, "John believes that that horse is brown." Such a belief relates John to a proposition, which, on Russell's (sometime) view, is a structured collection of individuals and properties. If we assume that demonstratives can be used to refer to concrete physical objects—not merely to sense data—the concrete horse is itself part of the relevant proposition. If, further, we take the content of a belief to be the proposition to which the thinker is thereby related, the content of John's belief that that horse is brown contains the very horse in question. And then the step to HEC: on the assumption that the contents of a subject's thoughts are parts of her cognitive mind, John's mind contains the horse; thus his mind extends into the extraorganismic physical environment.[17]

As cut and dried as this matter may seem, one who takes a Russellian

[16] Burge, "Belief *De Re*," this JOURNAL, LXXIV, 6 (June 1977): 338–62.

[17] If the Russellian argument in the text supports HEC, the same considerations cannot also support a content externalism according to which content is determined by something outside the cognitive system. For if the Russellian considerations prove HEC, the "external" factors of interest are part of the cognitive system and thus are not external to it.

view of thought content is free to resist the seemingly straightforward inference to HEC. Consider Colin McGinn's externalism. Motivated partly by Russellian considerations, McGinn arrives at a view much like HEC, which he takes to be the honest extension of content externalism.[18] It is, however, a mistake to identify McGinn's externalism with HEC. Granted, on McGinn's view, externalism implies that at least some of a thought's constituents exist beyond the physical boundary of the organism. Nevertheless, McGinn's conception of a thought's constituents differs significantly from the HEC theorist's conception. For McGinn, having a thought with a certain externalist content is for the substance of the person, that is, the body, to bear a certain relation to elements beyond the body; it is not a matter of the external elements appearing as parts in a configuration of things that is identical to a state of the mind, that is, the state of a mental substance; for on McGinn's view, the mind is not a substance, not even a physical one (*op. cit.*, pp. 24–26, 46, 116, 210). Thus, although McGinn often describes mental states as having "worldly constituents," when this talk is properly understood, such an interpretation of externalism does not imply externally located mereological parts of the mind.[19] In fact, it is partly the HEC-style implications of content externalism, when wed to a substantialist conception of the mind, that drive McGinn to reject substantialism about the mind: "We may thus finally declare that the mind is not a substance, and this because of externalism" (*op. cit.*, p. 103). McGinn rejects substantialism about the mind largely because together with content externalism—which McGinn takes

[18] *Mental Content* (Malden, MA: Basil Blackwell, 1989), pp. 37–43. McGinn also considers the relation between a Fregean view of content and content externalism. In this spirit, one might claim that in order for a subject to entertain a given concept, she must be related to an abstract entity, a property, in terms of which the concept in question is individuated; this view implies a form of content externalism that, one might think, implies HEC. (Cf. Ned Block's remark on the sense in which even narrow content goes beyond the boundaries of the head, if such abstract objects as concepts count as being "outside" of the physical head; "Advertisement for a Semantics for Psychology," in French, Uehling, and Wettstein, eds., *Midwest Studies in Philosophy, Volume 10: Studies in the Philosophy of Mind* (Minneapolis: Minnesota UP, 1986), pp. 615–78, note 7.) For reasons that will become obvious below, this Frege-inspired externalism is even farther from HEC than is a Russellian view of demonstrative thought content. Thus, I concentrate on the latter in my contrastive discussion of content externalism and HEC.

[19] Cf. Philip Pettit, *The Common Mind: An Essay on Psychology, Society, and Politics* (New York: Oxford, 1993), pp. 31, 43; David Papineau, *Philosophical Naturalism* (Malden, MA: Blackwell, 1993), pp. 29, 89. In addition, Pettit advocates what he calls an "attitude-based" externalism, according to which one must enter into, or be prepared to enter into, certain relations with other thinkers in order that one have any mental states with determinate thought content (p. 191). Here we encounter another content-externalist view that does not entail HEC.

to be firmly established—it implies that the mind has mereological components with external physical location (*op. cit.*, p. 20–22).[20]

It should now be clear that an externalist's talk of thought contents, their parts, and their constituents, can easily break free of the mereological domain; such talk is often not geared toward the placing of thoughts in their physical location. This should come as no surprise, for such externalist theses as McGinn's are set broadly within a tradition in philosophy of language that lacks the contemporary concern for (some might say "obsession with") naturalizing the mental[21]; this is not to say that McGinn opposes naturalism—to the contrary: construed broadly enough, he embraces it. All the same, HEC theorists typically pursue a different brand of naturalization, one taken to imply a much closer alignment between empirical work and issues traditionally discussed under the rubric 'philosophy of mind'; when advocates of HEC talk of cognitive states and cognitive processes, they often are concerned with the details of the physical realizations of cognitive systems—their physical arrangements and the changes in those physical arrangements, at least insofar as these seem to bear directly on cognitive processing. From this perspective, the question "Where are mental states located in physical space?" seems far more pressing.

Here is a final reason to reject the close association of content externalism and HEC. Recall the sorts of example externalists typically give in support of their views, examples where content—reference, most clearly—is determined by causal interaction between the subject and that to which the mental representation in question refers: the subject's 'water' concept refers to H_2O because she has had the right sort of causal intercourse with samples of H_2O. This kind of example plays little, if any, motivating role in the literature on extended cognition, and for good reason: HEC is not offered as a theory of content, that is, as an attempt to explain the semantic properties of mental or linguistic representations; HEC theorists do not propose that H_2O is a proper part of the typical subject's concept 'water' as a way of explaining the semantic properties of an internal unit, the water concept. In the

[20] In this respect, Rowlands presents a misleading interpretation of McGinn's view, placing McGinn in a camp with HEC theorists: upon explaining McGinn's externalist view, Rowlands claims that externalism—of McGinn's strong variety—"entails that mental states are located, at least in part, outside the skins of organisms that possess them," and for Rowlands this location-talk concerns physical location (see pp. 44–45).

[21] Cf. Kripke, pp. 96–97; Putnam, *Reason, Truth, and History* (Cambridge: Harvard, 1981), pp. 45–48. Although his case is less clear-cut, Burge would also seem to be in the camp of those not so enamored of the naturalistic project—see, for example, "Mind-Body Causation and Explanatory Practice," in John Heil and Alfred Mele, eds., *Mental Causation* (New York: Oxford, 1993), pp. 97–120, especially p. 116.

typical case, HEC arises instead out of an interest in the components of mental processes, out of an interest in nuts-and-bolts accounts of how cognition proceeds.[22] In what follows, then, the reader will find no further discussion of content, meaning, or reference.

IV. OTTO KNOWS MY PHONE NUMBER?

In support of HEC, Clark and Chalmers offer the hypothetical case of Otto, a victim of Alzheimer's disease who uses a notebook in much the same way most people use their internal memories. Imagine that Otto would like to go to the Museum of Modern Art (MoMA). Otto looks in his notebook, for there he has recorded MoMA's location: it is on 53rd Street, just east of the Avenue of the Americas. Clark and Chalmers claim that, given the way Otto treats the information in his notebook, that is, given the functional role it plays in Otto's cognitive economy, Otto believes the museum is on 53rd Street, even before he looks it up. His disposition toward that piece of information, recorded in his notebook, does not differ significantly from the average New Yorker's disposition toward her nonoccurrent (that is, not currently active or present to consciousness), but explicitly encoded,

[22] For something of a borderline case, see David Houghton, "Mental Content and External Representations," *Philosophical Quarterly*, XLVII (1997): 159–77.

Here it may be instructive to consider reasons for listing Millikan's essays among the work of HEC theorists (see note 1). On one hand, Millikan's approach to mental content shares much with orthodox content externalism, although put in terms of her biologically oriented approach: "Beliefs themselves are functionally classified, are 'individuated', not directly by function but according to the special conditions corresponding to them that must be met in the world if it is to be possible for them to contribute to proper functioning of the larger system in a historically normal way" (p. 189). On the other hand, Millikan sometimes endorses positions much closer to HEC: "It is a very serious error to think of the subject of the study of psychology and ethology as a system spatially contained within the shell or skin of an organism" (p. 158). As a claim about the subject matter of psychology, this might seem to be a mere claim of epistemological dependence: psychological properties are determined by biological function, and, according to Millikan, biological functions are best understood by considering the broader system of which a given organism is a part. But Millikan has something further, more HEC-like in mind: "The animal itself, considered as a system of events, extends far out into the extrabodily environment" (p. 180). Here, though, notice the focus on the *animal*. We might agree that animals are extended, without its being at all clear what follows from this regarding the location of the animal's psychological or cognitive states. Even if content is determined relationally (à la content externalism) and the study of psychological functions is broad (because of epistemological dependence) *and* animal systems are extended, it may yet be the case that physical belief states whose function it is to map onto factors external to the animal have no proper parts beyond the boundary of the organism. Thus, Millikan's view would appear to demonstrate a way in which one might advocate an extended self (the entire animal self) without advocating an extended mind or cognitive system. Whether Millikan conceives of her view in this way is less clear.

belief that MoMA is on a certain block of 53rd Street. In such cases, Clark and Chalmers claim, the "belief is simply not in the head" (*op. cit.*, p. 14).

The case of Otto's "belief" regarding MoMA's location reveals the flavor of Clark and Chalmers's reasoning: Otto's externally stored "belief" plays the same functional role in his cognitive system as do the typical person's internally stored, but nonoccurrent, beliefs.[23] But under what conditions exactly does cognition extend beyond the traditional subject? Clark and Chalmers list four general grounds for ascribing an extended belief to Otto:

> First, the notebook is a constant in Otto's life—in cases where the information in the notebook would be relevant, he will rarely take action without consulting it. Second, the information in the notebook is directly available without difficulty. Third, upon retrieving information from the notebook he automatically endorses it. Fourth, the information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement (*op. cit.*, p. 17).

Further consideration of the fourth criterion, the "past-endorsement criterion," as Clark and Chalmers sometimes call it, creates a dilemma for the HEC theorist. The HEC theorist has good reason to want to embrace the past-endorsement criterion, for the first three can be satisfied far too easily. Yet, the past-endorsement criterion runs counter to the spirit, if not the letter, of HEC: adopting the past-endorsement criterion undercuts HEC's motivation, by offering a picture of cognition that seems to fit better into the explanatory framework offered by HEMC.

An example will help to develop this dilemma. Prevalent in modern society are telephones, including cellular telephones, and a system of directory service. Given these facts, the first three criteria imply that virtually every adult, Otto included, with access to a telephone and directory service has true beliefs about the phone numbers of everyone whose number is listed. The directory assistance operator is a constant in Otto's life, easily reached; when the information would be relevant, it guides Otto's behavior[24]; and Otto automatically

[23] Cf. Donald's functionalist argument in support of HEC as applied to memory (p. 309).

[24] I have included this clause so that Otto's "belief"-states satisfy the first of Clark and Chalmers's criteria; yet, although the past-endorsement criterion is the primary cause of concern in the present section, this first criterion causes mischief as well, for it demands too much of a state in order for the state to count as a belief. Subjects often fail to consider internally stored beliefs that would be relevant in a situation at hand. Clark and Chalmers might require only that *when* the belief is active or has been accessed, it (almost always) guides the subject's behavior. They must, however, give an account of what it is to be active (or accessed) that does not privilege internal

endorses whatever the operator tells him, about phone numbers, anyway. It is absurd to say that Otto has beliefs about all of the phone numbers available to him through directory assistance (that is, beliefs of the form, "John Doe's phone number is ###-####"),[25] so long as he remembers how to dial up the operator. To say so would be to depart radically from the ordinary use of 'belief' (similar remarks apply to 'know': given ordinary usage, we would not say that Otto knows my phone number to be such-and-such). Inclusion of the past-endorsement criterion seems well advised, then. Choose at random a person with listed phone number such-and-such; assume, plausibly enough, that Otto has never consciously entertained the idea that that person's number is such-and-such; the past-endorsement criterion saves the HEC theorist the embarrassment of having to say that Otto has an accurate belief that the person's phone number is such-and-such.[26]

---

consciousness; if they cannot give such an account, the amended criterion causes a problem of internal privilege similar to the one—to be discussed below—caused by the past-endorsement criterion.

[25] The beliefs at issue are not merely beliefs about many unfamiliar items subsumed under a manageable description, for example, the belief that every number in the phone book was put there by employees of the phone company. Such examples raise difficult questions about dispositional, implicit, and *de dicto* belief. These are not, however, the questions of present concern; the HEC theorist claims that the subject's mind contains an explicit representation of the specific belief content in question, as is supposed to be the case with the written text in Otto's notebook.

[26] Clark considers an example similar to the one just described—*Being There*, p. 217—without seeming to appreciate the trouble it causes for HEC, even though Clark's criteria for extended states parallel those on the list presented by Clark and Chalmers. To be fair, Clark does offer the following further criterion, which might seem to solve the present problem: an external artifact must be "personally 'tailored'" for a subject in order that a state of the artifact constitute part of one of that subject's mental states. Nevertheless, Clark does not make clear what form of personal tailoring does not entail conscious endorsement, yet will serve the HEC theorist's purposes. If Otto were king and were to command that his subjects put together a telephone directory for his personal use, would that suffice as a form of personal tailoring? One should hope not.

In conversation, William Lycan has suggested a plausible, further criterion the HEC theorist might add in order to handle such difficulties as the one I have raised in the text (but note, Lycan's suggesting this defense of HEC should not be taken to imply that he endorses HEC): independent of questions about conscious awareness, the HEC theorist might simply require that the internal subject be *causally responsible* for the creation of the external marks or patterns that serve as memory traces in the external store. It is difficult to see, though, how this condition will inoculate HEC against the relevant problem-cases, unless the HEC theorist can appeal also to a criterion of conscious endorsement. At the very least, the HEC theorist seems forced to require that the subject have in some way grasped the meaning of what she is causally responsible for encoding. Otherwise, the subject might be causally responsible for creating information-expressing marks without plausibly standing in the belief-relation to the information expressed by these marks. Under the right circumstances, one can, for instance, create an enormous database by the mere stroke of a key, and these could be circumstances in which Clark and Chalmers's first three criteria are satisfied (as, it would seem, could be Clark's criterion of

On the other hand, at least the two following considerations speak against the HEC theorist's acceptance of the past-endorsement criterion. Clark and Chalmers, who have an ambivalent attitude toward the past-endorsement criterion, suggest the first (*op. cit.*, p. 17): a person can acquire ordinary, nonextended beliefs through processes of which she is not consciously aware; since it would be arbitrary to make the past-endorsement criterion necessary for extended belief, but not nonextended belief, it is best to give up said criterion. Secondly, adopting the past-endorsement criterion undermines what is supposed to be, if one accepts HEC theorists' revolutionary sounding rhetoric, one of the most important theoretical implications of HEC: that there is no good reason to assign special status to the boundary between organism and environment. If an extended (or any) belief requires conscious endorsement in order to be a genuinely held belief, and conscious endorsement is ultimately an internal process (that is, one that takes place within the organismic boundary),[27] then the traditional subject is privileged in a deep sense, after all.

One might wonder, however, about the seriousness of this last concern.

---

personal tailoring—the data might have been compiled off the Internet by a search program that takes personalized inputs). On the view under consideration, this would put the database-creating subject in a position, relative to the data in question, analogous to Otto's position with respect to numbers in the telephone directory. If, however, to head off this problem, the HEC theorist adds the requirement that the subject grasp the data's meaning, she seems to have re-adopted a version of the conscious-endorsement criterion, bringing in its train the attendant problems; furthermore, even if the HEC theorist were to require only an internal but *subconscious* grasp of the relevant meanings, the problem of internal privilege, raised below, would yet apply.

[27] Might conscious awareness itself be a property of an extended system? To transform this suggestion into a defense, the HEC theorist must develop an extended theory of *conscious acts* (a project from which Clark explicitly distances himself (*Being There*, pp. 215–16); for his part, Rowlands suggests that we dissolve the problem of consciousness (p. 2). An extended theory of conscious acts would be hard pressed to avoid an objection analogous to the one presented in the text, now constructed so as to apply to the acts of conscious endorsement rather than to nonoccurrent beliefs: if the act of conscious endorsement is made partly external, the HEC theorist would seem committed to saying that a subject has consciously endorsed data on such exiguous grounds as that the subject would, under certain circumstances, be inclined to accept the data—no occurrent awareness, in the traditional sense, required.

Note also the difficulty the present concern creates for a theory of socially distributed cognition. Would we be willing to attribute beliefs to a corporate body simply because it or one of its members has the sort of access to information that the average person has to the telephone directory? If not, how will the proponent of socially distributed cognition explain why not? Will she appeal to a conscious endorsement criterion? Such a tack does not seem promising: we have neither a plausible theory of consciousness for corporate bodies nor paradigm cases of conscious acts of corporate bodies.

As HEC has been defined here, it requires only that some mereological parts of cognitive states or processes be externally located. Since this demand is consistent with the sort of internal privilege embodied in the conscious-endorsement criterion, the HEC theorist might simply grasp the dilemma's second horn: she can accept the conscious-endorsement criterion and the accompanying internal privilege (perhaps softening her rhetoric a bit as a result).[28]

Although this response seems well-enough placed dialectically, it does little to obtund the damage done by HEC's welcoming of internal privilege:[29] if a subject's "external" memory or belief-content must be endorsed by organismically internal consciousness, it becomes more difficult to motivate the choice of HEC over HEMC; there is less reason to view external marks and objects as anything more than tools used by the mind, as opposed to parts of it. We can grant that cognition often involves intimate interaction with its environment. But given that internal consciousness provides the ultimate source of cognitive authority, it seems quite natural to say that the thinking subject, traditionally conceived of, is *using* those external resources. This way of putting matters, however, is best accommodated by HEMC; and given the costs to intuition—and to the general principle of conservatism in theory acceptance—of spreading the mind out into the world beyond the organism, there seems no reason to reinterpret the situation in keeping with HEC.

V. EXPLANATION AND COGNITIVE SCIENCE

In the detailed form given to it by Clark and Chalmers, HEC either implies highly counterintuitive attributions of belief or maintains an

[28] Herbert Simon argues for what might be interpreted as a version of HEC that retains internal privilege. Simon places external data storage on par with internal storage, for he claims that the structure of the external environment plays the same role as internal, long-term memory. Here sounding like a HEC theorist, Simon denies the significance, in at least one cognitive context, of the distinction between what is external to the organism and what is internal—see *The Sciences of the Artificial*, second edition (Cambridge: MIT, 1981), pp. 104, 117. At the same time, Simon maintains a privileged internal mind, in that he includes only internal processes as the fundamental processes of the mind; cognition results when these internal processes, such as methods of search through a database, interact with the environment, where the environment is construed in Simon's sense so as to include internal and external long-term memory. Simon's views are of great interest and may have inspired some advocates of HEC; nevertheless, Simon's thesis would seem much less radical than HEC, for rather than including part of the external environment in the mind, he places outside the mind an internal component that we normally take to be one of the mind's proper parts: internal, long-term memory (cf. Haugeland, p. 6; McClamrock, p. 89).

[29] Furthermore, notice that the HEC theorist's grasping the second horn of the dilemma does nothing to assuage Clark and Chalmers's own qualm about requiring conscious endorsement.

internal privilege that threatens to undermine the choice of HEC over HEMC. To a great extent, though, intuitions drove this conclusion: it is counterintuitive to attribute to Otto such extensive beliefs about or knowledge of phone numbers; and, given internal privilege, it seems more natural to stick with HEMC, rather than HEC (although here a principle of conservatism also played an important role). Such criticisms of HEC will likely be lost on those who do not share the intuitions bruited; stalemate looms. How can such disagreement be resolved? Is our choice of a theory of cognition merely a matter of "picture-preference"? A candidate arbiter waits in the wings, one suggested by the HEC theorist herself: a criterion of empirical fruitfulness. Whether in reaction to common-sense based objections to HEC or as an attempt to motivate HEC beyond mere HEMC, the HEC theorist can insist that she offers HEC as an explanatory hypothesis in cognitive science, and that HEC must be judged accordingly. Clark and Chalmers defend HEC in this manner:

> We do not intend to debate what is standard usage [of 'belief']; our broader point is that the notion of belief *ought* to be used so that Otto qualifies as having the belief in question.... By using the 'belief' notion in a wider way, it picks out something more akin to a natural kind. The notion becomes deeper and more unified, and is more useful in explanation (*op. cit.*, p. 14).

Rowlands takes a similar position. When introducing his argument for extended memory, he says, "I shall argue that, at least with regard to the memory systems possessed by modern human beings, there is *no sound theoretical reason* for setting up a dichotomy between internal memory processes and external aids to those processes" (*op. cit.*, p. 121, emphasis added).

The central thread of the argument would seem to be this: a taxonomy that includes overarching cognitive kinds[30]—kinds that cut across

[30] Here and in what follows I talk generally about explanatory and causal-explanatory kinds without having in mind too narrow a conception of such kinds. Broadly speaking, causal-explanatory kinds are those that support successful induction and explanatory practice in everyday life and, more to the point, the sciences. (Cf. W.V. Quine, "Natural Kinds," in *Ontological Relativity and Other Essays* (New York: Columbia, 1969), pp. 114–38; Philip Kitcher, "Species," *Philosophy of Science*, LI (1984): 308–33, note 11; Jerry Fodor, "Special Sciences," *Synthese*, XXVIII (1974): 77–115.) Although Clark and Chalmers use the term 'natural kind', I avoid doing so, for the following reason: if one conceives of natural kinds as the kinds to which natural kind terms refer, and one holds that natural kind terms refer to kinds the members or samples of which share microstructural essences (in the fashion advocated by Kripke; Putnam (*op. cit.*); and Putnam, "Is Semantics Possible?" in *Mind, Language, and Reality: Philosophical Papers*, Volume 2 (New York: Cambridge, 1975), pp. 139–52) then the HEC theorist's naturalistic gambit considered in the text seems open to easy and hollow refutation: since the external portions of the allegedly extended states almost certainly

the organism's boundary—provides the most empirically powerful framework for research in cognitive science. Since HEC, but not its competitors, offers the philosophical underpinnings for this framework, said gain in empirical power validates HEC.

This line of reasoning faces serious difficulties, as I intend to make clear in what follows. My strategy is to focus on a specific kind of cognitive state, memory, and here the thrust of the discussion is twofold: I argue that the external portions of extended "memory" states (processes) differ so greatly from internal memories (the process of remembering) that they should be treated as distinct kinds; this quells any temptation to argue for HEC from brute analogy (namely, extended cognitive states are like wholly internal ones; therefore, they are of the same explanatory cognitive kind; therefore, there are extended cognitive states). I argue further that the positing of a weakly defined type, *generic memory*, does not improve HEC's prospects. Although such a kind would clearly subsume some extended states, there is little, if any, causal-explanatory work for such a watered-down kind to do.

The value of the discussion to come might seem clear enough, then: HEC theorists often appeal to the case of memory to support HEC, and I will, in effect, be giving those arguments a critical going over. As worthwhile as such work may be, the discussion also supports broader conclusions pertaining to the overall evaluation of HEC. First, keep in mind that some HEC theorists claim to provide a radical reorientation to working cognitive science. John Haugeland, for example, advocates a reconception of intelligence that, if accepted, will change cognitive science root and branch (*op. cit.*, pp. 34–36). Interpreted in this way, HEC's plausibility depends on the widespread empirical success of HEC's taxonomy of cognitive states and processes; this taxonomy must provide a coherent and fruitful framework within which to place all, or at least a healthy majority of, significant results in cognitive science. Thus, HEC's failure to accommodate a wide range of results on memory constitutes a genuine strike against HEC. Second, bear in mind that memory is a fundamental cognitive process, subserving virtually all other important cognitive functions, including

---

do not to share microstructural essences with the portions of the human brain that realize or instantiate standard, nonextended mental states (and to which our rigidly designating mental-state-cum-natural-kind terms might be thought to refer), one could move straight to the conclusion that HEC is false. I do not do so. Instead I take explanatory practice in cognitive science at face value, without theoretical gloss. To the extent that one might interpret this practice in a way that is favorable to HEC, it would most likely be along functionalist lines; this approach is considered below, in section VIII.

language use and the storage of nonoccurrent beliefs—such as Otto's allegedly extended belief about MoMA; thus, it would seem unreasonably limiting to take the following discussion only to be a rebuke of HEC as applied in one narrow domain of cognitive studies, with no ramifications for HEC's prospects in other areas of cognitive research. Third, if any claim to taxonomical superiority stands a chance of providing significant support for HEC (and for an extended mind), it is a fairly strong claim to taxonomical superiority. Less sweeping claims, while not automatically defeated by HEMC, face a weighty burden of proof. If casting explanations in terms of overarching cognitive kinds provides only occasional benefit in cognitive science, the argument from taxonomical superiority will not do significant work in support of HEC; HEMC will suffice. Lastly, the discussion of memory is meant to serve as an object lesson, a way of highlighting the kinds of hurdle HEC must overcome in order to assume a foundational role in cognitive science. On, then, to a detailed discussion of memory.

Consider first Rowlands's argument (*op. cit.*, p. 133ff.), inspired largely by the work of Merlin Donald (*op. cit.*, passim pp. 308–33), for a general conception of memory as extended. Rowlands claims that as external stores—repositories of written language, for example—become widely used, memory strategies change: subjects begin to rely more heavily on external stores. As a result, states of external stores assume an indispensable information-bearing role in the process of remembering; and when this occurs, as it has in the case of modern humans, the relevant states of external stores become proper parts of the cognitive process of remembering (or of the states the transformation of which constitutes that process).

We should grant that acquiring the ability to write down, and later read, the contents of, for example, a speech, will be accompanied by a change in the structure of the relevant internal memory-related processing (although whether this is a change in mental *architecture*, as Donald claims (*op. cit.*, p. 273), is another matter). The subject no longer must work her internal episodic memory so hard, for she need not commit to internal memory the details of the speech. When she wants access to the contents of the speech, she need only read over a written version of it. In a society where code use is common, internal episodic memory may weaken among the populace, perhaps because of a kind of atrophy or the lack of a need to develop effective techniques for internally storing the details of particular experiences. Rowlands, however, does not make clear why the use of an internally represented code applied to the contents of an external store implies HEC, rather than HEMC. Although increased use of external resources might change the character of internal processing and the way in

which the subject interacts with her environment, why think that the apposite external and internal states (or forms of processing) are thereby of the same causal-explanatory kind? Why infer the existence of one overarching kind, memory, subsuming both internal and external states and processes, that will be of significant explanatory use in cognitive science? The HEC theorist might define 'memory' in a general way, thereby creating a category that includes external stores: consider Donald's proposal that memory is "a storage and retrieval system that allows humans to accumulate experience and knowledge" (*op. cit.*, p. 309). Characterized in such broad terms, the kind *memory* surely subsumes at least some external stores humans regularly exploit, but the HEC theorist cannot make such stores parts of human memory in one act of definition. The HEC theorist must motivate such a broad definition of 'memory' by putting the definition to work, by showing how the definition sheds otherwise unattainable light on established results or by running new experiments to demonstrate the value of HEC's framework of state types. A survey of the existing memory literature, though, should dampen the HEC theorist's enthusiasm for this strategy.

A wealth of memory-related research has focused on working memory—normally thought of as the especially active or accessible part of our internal memory resources. Rowlands claims that it is wrongheaded, from a causal-explanatory standpoint, to characterize working memory as an internal store; instead, Rowlands claims, working memory is "hybrid," a conglomeration of both internal and external stores, plus the processes that operate on these stores (*op. cit.*, pp. 145–46). Rowlands points out that internal working memory exhibits striking limitations, claiming that we can use it to carry out only the "simplest memory tasks" (*op. cit.*, p. 146)[31]; here Rowlands directs us to George Miller's classic work showing that humans can hold only a small number of items—approximately seven—in short-term memory.[32] This forces Rowlands to look elsewhere for resources that, working in tandem with severely limited internal processing capacities, explain how humans can quickly carry out complex, information-hungry cognitive tasks. Rowlands settles on a view that makes the external store primarily constitutive of working memory.[33]

Although we must recognize limitations on the capacity of working

[31] Also see Donald, pp. 328–29.

[32] "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, LXIII (1956): 81–97.

[33] "It is the information contained in these external structures which is the true *locus* of working memory" (p. 146).

memory, Rowlands's view fits poorly with much of the empirical data. Consider the human ability to converse effectively. Participating in a conversation of any significant length makes rigorous demands on working memory by requiring participants to build and maintain a fairly detailed model of the ongoing discourse.[34] If we take the HEC theorist at his word, there is no reason to assign a distinct role to internal storage in carrying out the conversation-facilitating tasks of memory; internal and external memory are simply two instantiations of one explanatory kind: working memory. This is, however, dead wrong. In the context of a standard verbal exchange of any significant length, external resources are virtually useless, while internal storage appears to be irreplaceable. Imagine that, in order to maintain a running account of an ongoing conversation, someone attempts to use one of the HEC theorist's favorite examples of an external store: written language. I, for one, do not relish the prospect of conversing with someone who maintains, on paper, a running model of our discourse. This is not a matter of my being impatient. An interlocutor's having continually to create and consult a record of the ongoing discussion, frantically writing and flipping through pages in her notebook to find the relevant parts, would destroy the dynamics of normal conversation.[35] The use of internal storage is the difference between successful and unsuccessful verbal interaction with other subjects, and thus a difference that a cognitive theory of conversation had better explain. The explanation will appeal, in the main, to the differing storage strategies being used; that such differences play an important explanatory role implies the presence of two different explanatory kinds.

Rowlands's dim view of the capacity of working memory seems to result from a limited focus on certain facts about the internal short-term memory store (number of stimulus items held: approximately seven; length of time unrehearsed stimulus stays in the phonological loop: less than two seconds). If, however, we broaden our view and

[34] See, for example, Connie Dickinson and T. Givón, "Memory and Conversation," in Givón, ed., *Conversation* (Philadelphia: J. Benjamins, 1997), pp. 91–132.

[35] Neither here nor in what follows do I mean to make the following complaint against HEC: the advocates of HEC cannot explain how an extended system that in fact gains access to relevant information in a timely fashion does so. Such a criticism would be unfair given that non-HEC-based cognitive science has been unable to explain how the individual gains access to relevant information in a timely fashion (thus the celebrated frame problem). Rather, I mean to criticize HEC for various forms of *mismatch* between memory as we know it in the standard case and what is alleged to be extended memory. This mismatch undermines the argument that external storage is enough like standard memory that the former should count as being of the kind *memory*.

appreciate the role of short-term memory within a larger system of working memory, internal working memory appears not so feeble a resource.

Short-term memory (STM) consists in "the retention of small amounts of material over brief time intervals."[36] In contrast, discussions of working memory often emphasize the use to which it is put: working memory is seen as part of "an integrated system for holding and manipulating information during the performance of complex cognitive tasks" (*ibid.*, p. 78). Working memory is taken to include not just a buffer (or multiple, modality-specific buffers) in which particular pieces of information remain "active," but also an executive that manages the buffer's (or buffers') resources. These differing descriptions of STM and working memory allow the two to come apart (*ibid.*, p. 83). Miller famously reported limitations on the number of items in an individual short-term memory store; these results, however, tell us little about how a central executive might manage the resources provided by various stores, how the information in an individual buffer is put to use for the purpose of carrying out complex tasks.[37]

The gap between STM and working memory becomes especially clear when one considers K. Anders Ericsson and Walter Kintsch's[38] hypothesis of a long-term working memory: a form of working memory that allows highly efficient access to more complex structures stored in internal *long-term memory* (LTM) (even in cases where such structures were only recently entered into LTM and may not remain stored "permanently").[39] This improved access to LTM is made possible by the maintenance, in STM, of a small number of cues that refer to parts of larger structures stored in LTM; these cued parts of the relevant LTM-structures are connected to the structures' other elements by, for example, semantic relations. The use of long-term working memory allows, for instance, access to a model of the content of a work being read, giving the subject ready access to material needed to disambiguate new portions of text as the subject reads them (*ibid.*, p. 230). Ericsson and Kintsch's theory explains how it might be that

[36] Alan Baddeley, "Short-Term and Working Memory," in Endel Tulving and Fergus I.M. Craik, eds., *The Oxford Handbook of Memory* (New York: Oxford, 2000), pp. 77–92, the quoted passage appears on p. 77.

[37] For discussion of the differences between the role, in language-processing, of a short-term memory store—the phonological loop, in particular—and working memory's central executive, see Susan E. Gathercole and Baddeley, *Working Memory and Language* (Hillsdale, NJ: Lawrence Erlbaum, 1993), chapter 8.

[38] "Long-Term Working Memory," *Psychological Review*, CII (1995): 211–45.

[39] For comparisons of access times to material in STM, LTM, and long-term working memory, see Ericsson and Kintsch, pp. 212, 215, 217, 224–25.

a system of internal working memory, with short-term memory buffers limited in the way Rowlands rightly observes, can make immediately available to the subject a large amount of information; and it can do so without, so to speak, going external.[40]

The human ability to converse draws heavily on what appears to be a complex system of working memory, and there is no reason to think that an external store can be effectively substituted for whatever component of this system serves to maintain an ongoing model of a conversation. In terms of Ericsson and Kintsch's model, conversation is made possible by a relation between a set of cues in internal STM and an internal model, held in LTM, of the conversation.[41] We have no reason to think that there is ever established a relevantly similar relation between a collection of cues in internal STM and an *externalized* model of a lengthy preceding conversation; and clearly little conversation will take place if both the retrieval cues and the model are outside the organism.

The HEC theorist might instead try placing the retrieval structure in the external environment, with the model of the ongoing conversation held internally; but what might serve as the external structure of cues?

[40] Miller himself illustrates the way in which chunking can greatly increase the amount of information held active in STM; his example involves the recoding of binary digits into orthographically simpler form, showing how a wealth of information can be packed into a small number of items held in STM (pp. 93–95). One should also keep in mind the extent to which elaboration or other forms of semantic or "deep" processing can increase the amount of information stored in memory; see John R. Anderson, *Learning and Memory*, second edition (New York: Wiley, 2000), pp. 198–202; and Scott C. Brown and Craik, "Encoding and Retrieval of Information," in Tulving and Craik, eds, pp. 93–107. Elaboration enhances performance on long-term memory tests by creating meaningful relations between elements. Cf. Ericsson and Kintsch where they review various elaborative mnemonic tricks used to store large amounts of information in long-term working memory (pp. 232–38). Also suggestive here is Gathercole and Baddeley's discussion of the drawing of inferences from text as a form of elaboration that improves children's comprehension of text (p. 228); this form of deep processing would seem to facilitate children's maintenance of models of text that are much like the running models maintained during conversation.

[41] I do not mean to lean too heavily on Ericsson and Kintsch's specific model of long-term working memory; there are, however, certain facts about cognitive processing that must be explained, among them the way in which participants in a conversation seem to keep large amounts of information at their fingertips without the use of external props (*pace* Donald, p. 343). If Ericsson and Kintsch's model does not accommodate these facts, they will have to be explained by some other model, which will either expand the powers of STM or explain how some information in LTM can remain highly accessible without props external to the organism (cf. Ericsson and Kintsch, p. 230). And note, the facts go well beyond language use: consider that chess grand masters can maintain accurate running models of numerous games—up to thirty at a time—blindfolded, that is, with no external props whatsoever (pp. 237–38).

Following Rowlands's lead (*op. cit.*, pp. 139–42),[42] the HEC theorist might take one participant's verbal production at a given time to provide the external cue-structure for the other participant. This, however, seems a hopeless suggestion. Imagine I issue a verbal rejoinder to an interlocutor's detailed criticism of a position I have stated earlier in our conversation. My interlocutor responds with the exclamation "Oh, pshaw!" On what plausible story could that bit of externalized sound provide a cue-set adequate for my going on in the conversation, thinking more carefully about the way my interlocutor had put his criticism and attempting to formulate my rejoinder in a more convincing manner? Donald's suggestion that conversations consist in the "recycling of common sentential utterances" (*op. cit.*, p. 370) might be helpful if we were out to explain a short exchange of pleasantries, but it is not a remotely plausible account of the lengthy and detailed conversation in which humans often engage. Furthermore, except in the context of a *completely* stereotyped conversation, where each person can respond in knee-jerk fashion to the last remark made, even a lengthy exchange of common utterances demands great memory capacity; only by the use of such a capacity can the participants produce *appropriate* common utterances—that is, only then do the admittedly stereotyped sentences compose a coherent conversation, rather than a series of pairs of associated comments.

Let us turn now to another kind of empirical result, not related specifically to short-term or working memory, that threatens the HEC theorist's causal-explanatory hypothesis: interference effects in paired-associates experiments. In paired-associates experiments, subjects learn assigned associations between pairs of stimulus items, with subjects' recall of these associations tested in various ways and at various time intervals. Negative transfer, a particular form of interference effect, appears when past learning detrimentally affects subjects' capacity to learn and remember new associations; it is observed in the following experimental paradigm, among others: experimenters direct subjects to memorize associations between pairs of words on a list—these might be names of men, as stimuli, and names of their female spouses, as the target responses. Call this first list of pairs the '*A-B*' list, *A*-words being those used as stimuli at the recall stage, *B*-words those that must be recalled upon exposure to *A*-words. The subjects learn, to criterion, the intended associations. In the next stage, experimenters shuffle

[42] Here Rowlands discusses not conversation, in particular, but the way in which the public act of verbalizing significantly improves recall of further related material, for example, the way in which reciting the opening lines of a poem makes it easier to remember the remaining lines.

the pairings, telling subjects, for example, that the couples in question have all divorced and remarried. Subjects are asked to learn the new pairs, on what is called the '*A-C*' list, and they do so significantly more slowly than they learned the *A-B* associations (or than they learn associations on a list made up of entirely new names). There is, it is said, negative transfer, an interference of the old associations with the learning of the new. The problem seems to be that if, for instance, John was married to Sally according to the *A-B* list, subjects have a hard time blocking out this association and forming a new association between 'John' and, say, 'Mary', with which 'John' is now paired on the *A-C* list.[43]

There is no reason to expect negative transfer in the learning of paired associates when a subject relies on an external store. The experimenter dictates the *A-B* list to the subject, and she records it in her notepad. After using the written list to answer the experimenter's questions, the subject sets it aside. Later the experimenter dictates the pairs on the *A-C* list to the subject, and she writes them down. Why would the items on the first list interfere with the accuracy of the data she enters on the second? The subject listens to the experimenter; she says, "John, Mary"; her words rebound through the subject's auditory working memory; the subject writes down the pair. Period. No problem, no interference. Similarly with recall: after the subject has recorded the *A-C* list, she sets it on the table for immediate access. When the experimenter provides only an A-word as stimulus, looking to the subject for the pair's completion, the subject simply consults her handwritten *A-C* list; presumably, she gets the right answer the first time, right away, with no negative transfer from related pairs on the *A-B* list. In fact, not only is there no interference; there is lacking entirely any typical learning curve for paired associates, under conditions that create interference or otherwise: assuming the subjects can take dictation and read their own handwriting, lists of pairs are "learned" immediately, on the first try, contrary to observations made under a wide variety of experimental conditions where subjects are allowed to use internal resources only. Granted, someone might lose her written list of paired associates, but there is no reason to think that, in general, "list-losing curves" will even approximate the forgetting curves found in paired-associates experiments.

[43] For descriptions of such experiments and further references, see Anderson, pp. 239–43; and Gordon H. Bower, "A Brief History of Memory Research," in Tulving and Craik, eds., pp. 3–32, especially pp. 9–14.

The above described differences between external and internal memory are neither trivial nor irrelevant from a cognitive standpoint. It is not a matter of saying, for example, that externally stored memories are typically a greater distance from my nose than are internally stored ones. The sort of difference to which I have drawn attention involves those characteristics—for example, learning time and access time—that are at the very heart of cognitive scientists' investigations of memory.[44] Furthermore, although I have illustrated the phenomenon of interference using the example of negative transfer in a paired-associates paradigm, this kind of effect appears in a wide range of cases. Gordon Bower describes thusly the pervasiveness of a kind of interference closely related to negative transfer:[45]

> [T]he basic ideas apply to analyses of forgetting in all learning situations such as serial learning, free recall, memorizing addition and multiplication tables, and remembering in which of multiple lists (or contexts) particular items occurred. It [*sic*] also applies to forgetting sentences, paragraphs, and stories when similar concepts are involved (*op. cit.*, pp. 13–14).

This indicates the enormity of the body of research that HEC theorists must account for—in the face of great difficulty, it would seem—if their causal-explanatory gambit is to succeed.

VI. HEC AND MALFUNCTION

Some proponents of HEC have recognized that, when faced with a malfunctioning extended cognitive system, it is useful to distinguish between the organism and the environment as separate components of that system.[46] Thus, the HEC theorist might claim that the internal/

[44] Compare the fundamental role Zenon Pylyshyn assigns to reaction times in the general investigation of cognition: according to Pylyshyn it is largely by measuring and comparing reaction tsimes that we can meaningfully identify and compare cognitive architectures; see *Computation and Cognition: Toward a Foundation for Cognitive Science* (Cambridge: MIT, 1984).

[45] Here Bower addresses the pervasive nature of *retroactive* interference effects, rather than negative transfer. In the case of retroactive interference, the learning of later material interferes with the subject's ability to recall earlier information. Matters are a bit less straightforward in the case of retroactive interference than in the case of negative transfer—see Michael J. Kahana, "Contingency Analyses of Memory," in Tulving and Craik, eds., pp. 59–72, here p. 62. Despite these complications, retroactive interference appears to be another phenomenon inexplicable from the standpoint of the HEC theorist who claims that memory conceived of generally, as either internal or external, is the conception of memory of greatest explanatory use to memory researchers.

[46] Clark, *Being There*, pp. 123–26; cf. Houghton, p. 171. Clark suggests this as part of his critical analysis of the dynamical systems hypothesis, which, among its other implications, is taken to support HEC—see van Gelder, pp. 373, 380.

external distinction has limited explanatory use in cognitive science—limited to the realm of cognitive breakdown. In respect of interference effects, the HEC theorist might claim, we should hardly be surprised that the internal/external distinction seems explanatorily relevant, for this is a case of breakdown, not a case in which memory functions properly.

Notice, however, that memory effects of the sort I have described are not limited to cases of malfunction; although this should be obvious from the discussion of conversation and working memory, the reader may be further persuaded by consideration of another well-confirmed memory-related phenomenon: the generation effect. The generation effect consists in a mnemonic *advantage* reaped by subjects who generate their own meaningful connections between pieces of material to be learned. An experiment run by Samuel A. Bobrow and Bower[47] provides an illustration. One group of subjects read sentences containing paired associates, for example, "The cow chased the ball" for the pair 'cow-ball', while another group generated their own sentences including the paired associates to be learned. The group that generated their own sentences performed significantly better than the read-only group when given a standard paired-associates completion test ('cow—??'). The generation effect exemplifies a robust characteristic of human memory: the general fact that elaborative processing, especially semantic processing, tends to improve performance on memory tests (for further references, see note 62). This is not a documentation of human failure, but a recipe for success.

Is there any reason to think that external memory stores exhibit the generation effect, that we can increase the strength of, or improve access to, what is stored in external memory by, say, having the extended portion of a system generate a connection between pairs, as opposed to its being fed the connection or being given no connection at all? What form would such externalist processing take? Notepads do not generate associations, at least not by themselves. Treating the organism-notebook unity as a single cognitive system, perhaps we should look for the generation effect to appear where external memory is used. Imagine the following experimental paradigm: in one condition, the experimenter enters paired associates, accompanied by context sentences of the experimenter's making, into subjects' notebooks. In the second condition, the experimenter enters paired associates into subjects' notebooks along with connecting sentences

[47] "Comprehension and Recall of Sentences," *Journal of Experimental Psychology*, LXXX (1969): 455–61.

that subjects themselves have generated. Given that we intend to test for a generation effect in extended "memory" systems, the experimenter will, during the testing period, provide subjects with unfettered access to their notebooks. There is no reason to think that subjects' test performance will vary depending on condition. Insofar as external storage drives subjects' responses, we should expect similar, possibly even identical, accuracy rates across conditions: regardless of condition, the subject simply looks in her notebook, sees the answer, and responds correctly—or, as the HEC theorist might have it, the organism-notebook unity emits an accurate response. If the test conditions were altered, so that subjects are not allowed to use their notebooks during testing, then one might expect to see a generation effect; but since subjects would not then be relying on external memory, such results would not show that external memory exhibits a generation effect.

In the set-up just described, the experimenter entered all data into subjects' notebooks. This arrangement focused our attention on two variables: place of storage (internal/external) and the generating source of context sentences (experimenter/subject). We did not, however, consider the *manner* in which the context sentences are entered into storage. Would it not be better to try to preserve the structure of Bobrow and Bower's experiment as much as possible, allowing subjects themselves to enter the data into their notebooks? After all, even when Bobrow and Bower fed context sentences to subjects, the information would seem to have, in some sense, passed through the subjects; they themselves entered the material into their long-term internal memories. Thus, it may seem more likely that extended memory would exhibit a generation effect in an experimental paradigm just like the one described above except that subjects do all of the writing.

On second thought, though, there is good reason not to expect such an outcome. Why would the process of having made up and entered a sentence make it any easier to find the correct page in a notebook than would the process of having merely entered a sentence? Perhaps if subjects' notebooks are cluttered, overstuffed, or, for some other reason, difficult to manage, self-entering might give subjects an advantage, but this would seem to hold as well for the context sentences generated by the experimenter. Furthermore, even if a generation effect appears and results from the use of external, as opposed to internal, resources, this would seem *entirely optional*, unlike the case of internal memory. If the subject so chooses, she can simply write *all* of the context sentences, generated and nongenerated, and all pairs on one page, rendering all of the pairs equally easy to

find (modulo scanning times). Clearly there would be no generation effect in this case.

The HEC theorist cannot rely on the distinction between properly functioning and malfunctioning systems to settle the concerns raised in the preceding section. In particular, the HEC theorist cannot treat instances of breakdown as collectively constituting the isolated, but principled, group of cases where we should expect HEC to forfeit its explanatory advantage; such an advantage is absent when it comes to the explanation of many memory-related phenomena that do not plausibly involve malfunction, including the generation effect and the use of working memory in conversation.

We seem forced, then, to recognize two different explanatory kinds, *internal memory* and *external resources used as memory aids*, with no reason yet found to think that external aids constitute genuinely cognitive states or processes. It might be possible to contrive external means of storage the use of which would exhibit characteristics parallel to those exhibited by subjects' use of internal memory. But even to concede this possibility (perhaps an ill-advised move) yet leaves a deep question unanswered: Why is it that particular learning curves, interference patterns, and so forth, are unavoidable when we rely on internal storage, while entirely optional, in fact, difficult to locate or produce, in the case of external storage? It seems that whatever combination of forces results in the cognitively relevant facts about humans' internal memory systems appears neither in the bits of organized matter one finds in the world external to the human organism nor in the relation between a human organism and those bits of matter.[48]

VII. GENERIC COGNITIVE KINDS

The HEC theorist might rejoin thusly: implicit in the preceding discussion is a common view about the metaphysics of kinds: the best indication that a given kind is genuine is that it plays a causal-explanatory role in our most successful science. And although the preceding considerations establish that internal and external memory constitute different explanatory kinds at some level (no surprise, perhaps, to any participant in the debate), we should not forget that taxonomies of kinds typically take hierarchical form. Perfectly consistent with the existence of two or more kinds of memory at lower levels in the

[48] Cf. "The resource managements techniques you are born with make no distinction between interior and exterior things"—Dennett, p. 142. The arguments in the text seem to show that internal resource-management techniques do make such a distinction, at least insofar as the use of different kinds of resource results in significantly different measurable effects.

hierarchy is the existence of an overarching kind, *generic memory*, such that, say, a subject remembers *P* if and only if the subject has some sort of access to the information that *P*.[49] Generic memory has explanatory utility, the HEC theorist might claim, and given that some extended states instantiate this kind, extended cognition is a reality; and insofar as this general kind of cognitive state underlies (or is identical to) the general mental state kind *remembering*, its instantiation would seem to entail an extended mind. The HEC theorist might bolster this claim by citing a trend in memory research toward the positing of multiple internal memory systems,[50] different from each other in important ways, yet grouped under the general rubric "memory." Given this development, the HEC theorist might wonder why we should not recognize external forms of memory as further memory systems, different in some ways from internal memory systems (and from each other, for that matter), but not any more so than various internal systems are from each other. Why should we not see all of these systems as instances of generic memory?

The appeal to multiple memory systems faces two significant problems: first, even if there is some diversity among internal memory systems, there is also evidence of substantial coherence among these systems—a coherence that renders them closer in kind to each other than any one is to what are typically claimed by HEC theorists to be external memory systems. For example, John Anderson observes that much of the data on remembering and forgetting—data collected in a variety of experimental paradigms—can be fit by power functions; thus, he postulates the power laws of learning and forgetting.[51] Anderson also argues for a general fit between the statistical properties of recall and the organism's needs given the statistical properties of the environment.[52] Consider also such learning functions as the Rescorla-Wagner law; although not of perfectly general application, it would seem to account for a wide range of data, rather than being tied to any particular setting or memory system.[53] There is little reason to assume that external stores exhibit these general features shared by many, perhaps most, internal memory systems.

[49] This response is suggested by Donald's discussion: he lists differences between internally and externally stored memory (p. 315) while also offering the general characterization quoted above, in section v.

[50] See Kahana.

[51] Anderson: on learning, see p. 187ff.; on forgetting, see p. 228ff.

[52] See, for example, his discussion of the spacing effect, pp. 238–39, and memory more generally in *The Adaptive Character of Thought* (Hillsdale, NJ: Lawrence Erlbaum, 1990), chapter 2.

[53] Anderson, (au: which title?) pp. 65–75.

Second, cognitive psychologists inclined to see memory as a fragmented kind are also inclined to reject the idea of a broadly defined kind *memory*.[54] Thus, the appeal to a proliferation of memory systems does not seem to do the work that the HEC theorist would like it to do. The existence of a variety of internal memory systems is supposed to show that a more general explanatory kind subsumes that variety. One might, however, just as well infer, as some cognitive psychologists have, that there is no such superordinate kind, *memory*.

In order to make the present line of argument convincing, the HEC theorist must establish that the mere fact of a state's being an instance of a broad cognitive kind (*generic memory*, *generic belief*) has explanatory value in a significant number of cases. Here the HEC theorist might appeal to explanatory simplicity. Consider Clark and Chalmers's argument in this regard: in place of HEC-style explanations, Clark and Chalmers point out that "one could always try to explain my action in terms of internal processes and a long series of 'inputs' and 'actions', but this explanation would be needlessly complex" (*op. cit.*, p. 10). Given the preceding discussion, though, we should doubt that such explanations will be *needlessly* complex. One expects instead that such complexity will be illuminating. A complex—either traditional or HEMC-style explanation—will shed genuine light, obscured by HEC, on the reasons for various differences in the way external and internal stores are accessed and used. Imagine that Otto uses his notebook to store paired associates, while Sarah stores them internally, leaving her notebook blank. The two systems, Otto-plus-notebook and Sarah-the-organism-alone, behave differently when quizzed. The interaction between Otto and his notebook together with the properties of the notebook itself explain why Otto displays different recall characteristics from those displayed by Sarah. Such an explanation will advert largely to the sort of interactive step—for example, inputs to Otto-the-organism—that Clark and Chalmers claim to be explanatorily gratuitous, yet there appears to be no equally powerful, but simpler, HEC-based explanation of the relevant behavioral differences.

The cognitively relevant properties of a piece of information's being accessible to the subject would seem to vary greatly depending on what store that unit of information resides in and how, as a result of its place of storage, the cognitive system gains access to it—so much so that it is not clear what useful role the kind *generic memory* plays in any real-life research program. Consider a further case, though, that

[54] Tulving, "Concepts of Memory," in Tulving and Craik, eds., pp. 33–43, p. 41.

might establish an explanatory role for generic memory as well as an analogous kind, generic knowing: a person who lives in the library knows, and perhaps remembers, everything in all books to which he has access, in a generic sense of 'knows' (ignoring for the moment that the possibility that HEC include the conscious endorsement criterion—inclusion of which, as we have seen, makes it difficult to motivate HEC over HEMC). What aspect of that person's behavior is both (a) complex and representation-hungry[55] enough to require cognitivist explanation, and (b) general enough such that its explanation depends only (or at least primarily) on the subject's general access to information, not on the fact that there is a particular kind of access (look-it-up-in-the-stacks access, as opposed to close-your-eyes-and-remember-it access)? Is it just that he correctly answers certain questions that some other people do not answer correctly? Given sufficient time and motivation, most people will find answers to difficult questions. Differences between cases in which people get right answers and those in which they get wrong ones depend, as much as anything, on the kind of access the organism has to the information in question and on the way in which the organism goes about trying to locate the information. If the general notion of access to information adds any explanatory power, it is too little to justify new ontological commitments; there is available a perfectly satisfying, HEMC-style explanation of the abilities of the man who lives in the library and how they contrast with the abilities of others, an explanation that invokes theoretical tools and commitments for which everyone must recognize an independently motivated need. Thus, even in cases where HEMC does not enjoy a clear advantage over HEC in other respects, a methodological principle of conservatism recommends HEMC over HEC.

VIII. FUNCTIONALISM AND HEC

Clark and Chalmers do not present their position as an explicit development of the functionalist program in philosophy of mind; nevertheless, the argument from Otto's case contains a clear functionalist strain. Driving the argument seems to be the idea that external encodings of information can play the same functional role as that played by internal encodings: the former might at least partly realize a mental state in the way internal encodings of information are often thought to. The HEC theorist might, then, generalize such considerations, formulating the following functionalist argument for HEC:

[55] The term is Clark and Toribio's—see "Doing without Representing?" *Synthese*, CI (1994): 401–31, p. 418, and *Being There*, p. 149.

Premise 1: A mental state of kind *F* is realized by whatever physical state plays the functional role that is characteristic (or metaphysically individuative) of *F*.

Premise 2: Some realizations of functional mental state kinds have physical components external to the organism.

Premise 3: A mental state extends to or includes all components of its realization.

Therefore, some mental states extend beyond the boundaries of the organism.

The argument's form is unobjectionable. We should wonder, though, what sort of functionalist view stands the best chance of offering plausible support for HEC. In particular, we should want to know what justifies the formulation of a particular functionalist psychological theory—say, in the form of a Ramsey sentence[56]—that would allow for extended states; for the plausibility of premise #2 depends on the particular functionalist theory of mind on offer.

Consider first the functionalist approach according to which analysis of common-sense psychological concepts yields functional-role descriptions of mental or cognitive states. For example, a memory that *P* is, among other things, caused by interaction with a certain state of affairs (which we might normally describe as the content of *P* or what the memory that *P* is a memory of) and, under certain conditions, a cause of the belief that *P*. This approach fails miserably. The analysis of common-sense concepts of cognitive states does not support HEC, for common sense rules strongly against external portions of memories and other cognitive states. The common-sense conception of memory precludes its being *seen* by its possessor (that is, precludes its being a cause of a perceptual, as opposed to an imaginative, state). Applying a common method of identifying the entailments of common-sense concepts, one should test one's reaction to the sentence, "Yesterday I saw my memory of last week's trip to the beach." The sentence exudes semantic deviance, and thus one should strongly suspect that the literal seeing of one's memories (or even parts of them) does not square with the everyday concept of memory. Similarly, the common-sense functional characterization of belief precludes the encoding of Otto's belief states in his notebook (or in the phone book), for ac-

[56] See David Lewis, "How to Define Theoretical Terms," this JOURNAL, LXVII (July 9, 1970): 427–46, and "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy*, L (1972): 249–58; Block, "Introduction: What Is Functionalism," in Block, ed., *Readings in the Philosophy of Psychology*, Volume 1 (Cambridge: Harvard, 1980), pp. 171–84.

cording to the common-sense conception of belief that functionalist theory must capture, a subject cannot literally see portions of her belief states. The Ramsey sentence expressing folk psychological theory will not assert a causal (or even merely conditional) relation between belief states and perceptual states, where the latter is a visual perception of a portion of the former; if any statement regarding such a connection makes its way into the Ramsification of common-sense psychology, it will be an explicit denial of the connection.

It might be of little concern to the HEC theorist that our attempt to Ramsify common-sense psychology fails to support HEC. Good functionalist analysis, and philosophical theory more generally, results only from careful reflection on a variety of examples and considerations; only then can we separate the wheat from the chaff in the folk understanding of the relevant concepts (or so it might plausibly be claimed). Thus, the HEC theorist might continue, we should cast a careful eye on the folk's rejection of extended states and instead formulate our functionalist analysis following a more refined method. What, though, should motivate refinement in the present context? Given that HEC is typically offered as part of the philosophical foundations of cognitive science—that, at least, is the angle of interest here—empirical considerations should motivate our functionalist theory. We must have some good reason for deciding that certain functional-role properties will appear in our Ramsification of psychological theory and that others will not. Since the price of admission is empirically productive service, the functionalism at issue should be a form of psychofunctionalism: a theory whose characterization of mental states' individuating functional roles is given by our best psychological theories. Even if, suspicious of baldly naturalist arguments, one conceives of psychological theory as an abstract analytical tool, it is meant to be an analytic tool that serves certain purposes: the explanation and prediction of behavior. Thus, even the rationalistic psychologist should give weight to the broadly empirical goals of explanatory and predictive success when choosing among possible psychological theories each of which is a candidate for Ramsification (or a candidate for a formal rendering of the intentional stance).[57]

The functionalist argument for HEC faces dim prospects as a psychofunctionalist argument. As cognitive science currently describes its explanatory kinds, they are not likely to have realizations with external components. If, for example, cognitive science is to characterize functionally the causal role of memories, this characterization

[57] See Dennett, *The Intentional Stance* (Cambridge: MIT, 1987).

must be tailored to accommodate the generation effect, various forms of interference, the power laws of learning and forgetting, and the rest. The resulting characterization speaks strongly against premise 2 of the functionalist argument for HEC. In response, the HEC theorist might attempt to show that it is of significant value to include in psychological theory such overarching kinds as *generic memory*. If motivation for including such kinds can be found, then perhaps the functionalist argument will go through, for the individuating causal role of such kinds is so broadly put that they are likely to have extended realizers. This alternative hardly seems more promising than the first, however. As argued above, whatever advantage, if any, derives from including such generic kinds in our taxonomy of mental states (now corresponding to the values of variables in a Ramsey sentence) would not appear to outweigh HEC's violation of conservatism, which instead encourages us to make do with HEMC. Thus, functionalism seems not to offer independent support for HEC: the functionalist-minded HEC theorist remains in need of an argument for including, in functionalist psychology, mental states or properties that are realized by or instantiated in extended states.[58]

[58] Further difficulties speak against a functionalist definition of 'generic memory'. Many states we would normally count as memories fail to exhibit reliably what would seem to be the defining functional aspects of memory. Internal memory states are not always readily accessible, do not always guide action even when their content would be relevant, and are not always treated as trustworthy by the subject of those states (cf. Clark and Chalmers's criteria for extended nonoccurrent beliefs). Many memories do not come when called and sometimes we simply "forget to remember"—the result being that we often act without considering relevant information stored in memory; and some memories are not treated as reliable by the agent that has them because, for example, they are accompanied by a feeling of uncertainty. Of greater importance than these functional traits, then, are causal-historical factors: an acceptable definition of 'memory' will place necessary conditions on the causal processes by which memories were stored. Donald's definition defers somewhat to causal history, in that it requires information in external memory to have been stored as the result of "experience"; however, given that, on Donald's view, the experience in question need not be that of the very subject who uses the external store, this requirement is too weak. It is anyone's guess, though, how the HEC theorist might strengthen Donald's requirement in such a way that it avoids both panpsychism (memories being widespread furniture of the universe) and the privileging of internal storage (as, for example, the necessary locus of changes resulting from the causal interactions requisite for memory formation). The fundamental difficulty here would seem to be that we need first to identify what counts as a cognitive system; this is the topic of section IX, below.

The HEC theorist who presses functionalism into service must also confront functionalism's shortcomings as a foundation for cognitive science, in particular, functionalism's difficulty explaining how cognitive states could be causally efficacious. See Block, "Can the Mind Change the World?" in George Boolos, ed., *Meaning and Method: Essays in Honor of Hilary Putnam* (New York: Cambridge, 1990), pp. 137–70; Jaegwon Kim, *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (Cambridge: MIT, 1998).

IX. THE PRIORITY OF COGNITIVE SYSTEMS

The preceding discussion has focused primarily on the question of how best to explain the cognitive and mental capacities of subjects traditionally conceived of. The HEC theorist might, however, claim that I have latched onto the wrong explananda for cognitive science. She might insist that our theorizing begin at an earlier stage, one at which we identify the truly cognitive systems (that is, systems the behavior or capacities of which are to be given cognitivist explanation). It might be that HEC gets its hold at this stage, for some extended systems might appear among those that possess cognitive or mental capacities. If they do, then so long as some extraorganismic parts of extended systems play a role in the best cognitivist explanation of those systems' behavior or capacities, the existence of extended cognitive states is sufficiently established.

We should not ignore the question of cognitive-systems delineation, but the HEC theorist must motivate the claim that extended systems possess truly cognitive or mental capacities.[59] We would like an account of stable cognitive *systems* such that we know when cognitive or mental states should be attributed to those systems. One *could* cook up an extended system, attribute a capacity to the whole system, then attribute cognitive states to the extended system in order to explain that capacity; and some of these "explanatory" states might have proper parts that exist beyond the boundaries of an organism at least some of which is part of the cooked-up system. Nevertheless, this would not show that the explananda of cognitive science comprise such systems' capacities: we want first to be convinced that such systems' capacities are the cognitive or mental capacities of integrated cognitive agents or thinking subjects.

[59] Hutchins attempts this in the closing chapter of *Cognition in the Wild*, arguing that mentality itself should be fundamentally reconceived, or perhaps more accurately, the proper conception of it must be rediscovered. Hutchins claims that clear conceptions of cognition and mentality—not those subverted by contemporary cognitive science—apply to extended systems as paradigm cases. Although such an argument, were it successful, would strengthen the HEC theorist's case, she should try to improve on Hutchins's offering; for Hutchins derives his conclusion largely from a questionable "deconstruction" of cognitive science, recounting, in the form of an anecdote, what is supposed to have been a foundational error in Turing's conception of his own abilities: Turing failed to see that the cognitive capacities he hoped to model computationally were capacities of extended systems, not of individual human organisms (pp. 360ff; for a similar point, put in more general terms, see Donald, p. 313). This line does not, however, account for what are obviously cognitive capacities of individuals, traditionally conceived of; take language use: when a subject formulates a grammatical sentence and speaks it out, no extended system produces the sentence—the organism does. (And for an objection to putting socially distributed cognition first, see note 38 above.) Of course, various external factors have helped to shape the organism's capacity to produce sentences, but to include all such influences as parts of a single, language-using cognitive system appears to be motivated by no more than the misguided principle of epistemological dependence criticized above, in section II.

In the end, empirical research should decide this question: we should commit resources to the framework of extended cognitive systems, apply the extended view in the study and the lab, and see whether doing so generates a flourishing research program in cognitive science. It is very difficult to predict the future of science; matters might work out in favor of extended systems. There are, however, reasons for pessimism. Let us consider two in closing.

One of cognitive science's most important undertakings has been the creation of artificial intelligence. What would become of A.I.'s research program if cognitive scientists were to think of extended systems as the paradigm possessors of cognitive capacities? Would the environment in which an A.I. system is to function be made part of that system? Would each project in A.I. involve the creation of an environment that travels along with the locus of computing? Researchers in A.I. are not, for example, much interested in creating a self-contained system—patient plus diagnosing A.I. program installed in a hardware module; rather, they wish to create an A.I. program to which we can present any given patient, receiving then an accurate diagnosis. Individuating cognitive systems in a broad way, so as to include the environments in which they function, would only seem to hinder A.I. projects.[60] Of course, the biases of A.I. researchers might be ignored by the HEC theorist (such researchers have yet to see the value of creating extended systems, it might be thought). All the same, as we know them now, the intelligence of A.I. systems consists largely in their flexibility as self-contained units that function effectively in various environments. In contrast, putting more of the environment into an A.I. system seems to make it less flexible, making it difficult to see what would be intelligent about such an extended system.

Consider a quite different but equally central component of cognitive science: developmental psychology. Although it has sometimes been claimed that the developing system should be conceived of as integrated with its environment,[61] this seems to have little plausibility when one takes a larger view of the purpose of developmental theorizing. Here the explanandum consists in a set of skills acquired by a system over the course of its development. The various cognitive skills to be explained are, it is to be emphasized, skills of a single coherent system, one with historical integrity; our theoretical account of the

[60] Even less traditional projects in A.I., Brooks's, for example, typically build discrete, self-contained systems that perform by interacting, HEMC-style, with their external environments.

[61] See Esther Thelen and Linda Smith, *A Dynamic Systems Approach to the Development of Cognition and Action* (Cambridge: MIT, 1994).

genesis of those skills should focus on *that very* system. In a typical developmental process, the environmental objects interacted with are dispensable and variable and thus do not seem to be parts of an integrated system that persists over time. Admittedly, a system's constitution can change over time while remaining the same system (compare Locke's examples of changing organisms that retain their identity in virtue of retaining their form and organization); noting this fact, the HEC theorist might claim that, even though the floor I learned to walk on as a child is not present now, something is present that plays a similar role in a single person-floor system. What criteria, though, would license the inclusion of the present floor, quite different from the developmental one, as part of the overall cognitive (or motor) system whose behavior is to be explained? A functionalist theory? On what grounds? We entered into discussion of systems-delineation partly because functionalist theorizing alone does not resolve the issue of extended states; there is no more reason to think it will resolve the issue of extended systems, for here again the functionalist must look to the empirical work to tell her which systems are to be subject to her psychofunctional theory (that is, to which systems her Ramsified theory must correctly apply).

The HEC-minded developmental theorist faces a dilemma. She must account for the enormous degree of flexibility present in the application of the skills that constitute developmental psychology's explananda. Such flexibility undermines the HEC theorist's attempt to describe at an empirically useful level of detail the development of extended systems. She must give an adequately elaborated account of the developmental process while being sure to describe only as much structure as will accommodate the "replacement" of external elements with their alleged functional or structural analogues in the persisting composite system. The former goal often requires detailed accounts of developmentally important interactions with specific external objects; the latter demands that one speak at a gross level. The best way to satisfy the latter goal (by saying, for instance, that children develop language skills by interacting with language in general) is to sacrifice the details of the developmental theory (children interact with *these* bits of language in *these* particular ways resulting in the acquisition of a skill that can then be exercised in a wide range of new cases). Better HEMC, then, and not just for the sake of parsimony; it allows us to articulate the important difference between what the developing subject gets from the objects with which she interacts and how she goes on to apply skills so acquired to quite different objects later in life: she represents certain aspects of those experiences (or the things experiences), combining and applying those representations

when handling new cases encountered later in life. This distinction would seem useful even if at every point in the exercise of some given skill, the (adult or juvenile) subject's performance is dependent on and greatly facilitated by the presence of a general kind of triggering stimulus (language) or object (the floor).

At this juncture, one might consider abandoning the attempt at uniquely cognitivist theoretical research, moving instead to the study of complex systems in general: individual human systems, ant colonies, whirlpools, and extended systems that include individual human organisms together with external elements, among other possibilities. This might be a viable route for a future science to take, but it is not consistent with HEC: within such an eliminativist framework, mind and cognition—extended or otherwise—no longer appear as causal-explanatory kinds. Eschewing these radical implications, the HEC theorist might still hope to secure a fundamental theoretical role for extended states or systems in the study of cognition. As things stand, though, HEMC provides the best interpretation of cases where intimate interaction between the organism and its environment supports what we normally take to be the cognitive and mental capacities of systems clearly in possession of them. The present state of affairs thus favors a healthy skepticism regarding the claim that HEC yields a more useful taxonomy—of states or systems—for the causal-explanatory purposes of cognitive science.[62] And without the strength of cognitive science's successes supporting it, the hypotheses of extended mind and self seem weak indeed.

ROBERT D. RUPERT

Texas Tech University

[62] HEC has recently come under fire from other quarters. Fred Adams and Ken Aizawa criticize and reject what is essentially HEC, although they use the label "transcranial cognition"—see "The Bounds of Cognition," *Philosophical Psychology*, XIV (2001): 43–64. A few words are in order, then, about the relation between my critique of HEC and Adams and Aizawa's criticisms of the hypothesis of transcranial cognition. Adams and Aizawa rest their criticisms largely on the distinction between derived and nonderived representation, an approach that I avoid entirely (without a thorough attempt to apply extant theories of intentional content to the allegedly external representations, the labeling of these as 'derived representations' seems to beg the question against the HEC theorist). Adams and Aizawa also argue that intracranial processes manifest different kinds from those found in allegedly cognitive, extracranial processes. Here they focus primarily on the *physical* differences between the intracranial and extracranial processes (pp. 46, 59), which seems at best to be only indirectly related to present concerns; more to the point, Adams and Aizawa sometimes worry that at the level of *cognitive* description, intracranial processes exhibit properties not shared by extracranial processes (pp. 61–62; also see a passing remark about *psychological* laws—p. 58). Although developed independently of Adams and Aizawa's work, some of what the reader finds in the latter sections of the present essay dovetails their worry that extracranial and intracranial cognitive processes exhibit distinctive, cognitively relevant properties.