"Memory, natural kinds, and cognitive extension; or, Martians don't remember, and cognitive science is not about cognition"[·]

Robert D. Rupert, U. of Colorado, Boulder

I. The natural-kinds argument for the extended mind

The last quarter of a century has seen an explosion of experimental and theoretical work in what might generally be called 'situated' cognitive science (Robbins and Aydede 2008). Such work emphasizes the role of the organism's (or artificial agent's) interaction with its environment during cognitive processing (that is, processing that produces intelligent behavior). Early contributions to the field can be found in Rodney Brooks's work in robotics (Brooks 1986), Esther Thelen and Linda Smith's research in developmental psychology (Thelen and Smith 1994), David Kirsh's exploration of epistemic actions (Kirsh and Maglio 1994, Kirsh 1995), Dana Ballard's research on animate vision (Ballard 1991), and Tim van Gelder's dynamicist philosophy of cognition (van Gelder 1995), among many other sources. Inspired partly by this burgeoning research program, Andy Clark and David Chalmers (1998) wrote a highly influential paper, titled simply "The Extended Mind." In it, Clark and Chalmers advanced the bold hypothesis that the realizers of human mental states and cognitive processes are (frequently?) at least partly located beyond the boundary of the organism. They claimed

---

[·] For comments on a previous draft of this paper, I extend my thanks to John Sutton, Kirk Michaelian, and an anonymous referee. For recent helpful conversations on the matters discussed herein, thanks to Ken Aizawa, Larry Shapiro, Alastair Norcross, and, especially, Mike Wheeler.

that such a view has sweeping implications for the "methodology of research in cognitive science" as well as for our thinking about "moral and social domains" (1998, 18). The paper closes in dramatic pronouncement: "once the hegemony of skin and skull is usurped, we may be able to see ourselves more truly as creatures of the world" (ibid).

What might justify such a radical vision of human thought? Clark and Chalmers present various supporting arguments, but one holds special interest for those oriented toward cognitive science: their argument from natural (or causal-explanatory) kinds. After making their case for extended beliefs, via the now-famous example of Otto the Alzheimer's patient who stores information in his notebook, Clark and Chalmers remark:

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. (1998, 14; also see their comments about explanation at 9–10)

How does the discussion of Otto's notebook bear on situated cognitive science? Here are three possibilities:

A. It provides a merely schematic model for genuinely naturalistic reasoning. The particular case may be somewhat far-fetched, and, yes, cognitive scientists may not seem

to focus much on belief or belief-desire explanations.[1] Nevertheless, the suggested strategy is clear: argue for extended cognition by trying to show that our best cognitive

[1] This should give Mark Sprevak pause. In his discussions of inference to the best scientific hypothesis, Sprevak (2009, 524; 2010, 358) claims that Clark and Chalmers's extended explanation of Otto's behavior has an advantage over an embedded one (Rupert 2004), by dint of its coarse-grained approach to action-explanation. But, so far as I can tell from my day-to-day involvement in the cognitive science unit at my institution, hardly a research program in cognitive science focuses on such coarse-grained belief-desire explanations as are deployed in Otto's case.

Readers not convinced of the marginalized position of coarse-grained, commonsense belief-desire explanation should consider the following argument (which might called 'the argument from nested models' – cf. Giere 2006, 717–718): When (a) a given *explanandum* can be successfully accounted for by both a less-articulated model and a more articulated one, (b) the less articulated model is amenable to alternative ontological readings, and (c) the more articulated model accounts for more of the variance in the relevant behavioral data, then we should derive our ontological conclusions (or the procedurally oriented equivalent) from the more articulated model. Our current situation satisfies the compound antecedent of the preceding conditional; Otto's behavior is subject to two explanations that involve equally coarse grained states: one is the belief-desire explanation offered by Clark and Chalmers (and endorsed by Sprevak), and the other involves a coarse-grained desire to go to the Museum of Modern Art (MoMA), a coarse-grained perceptual state of Otto's seeing his notes about the location of MoMA, and the resulting coarse-grained belief that MoMA is

science deals in natural kinds – either fine-grained or coarse-grained – that have both internally located and externally located instances, regardless of whether these kinds bear the names of, or are identical to, kinds discussed by the folk.

B. The discussion of Otto provides a model of how relatively pretheoretical conceptions of mental states interface with cognitive science. *A priori* reasoning about the coarse-grained causal roles of folk mental states, together with everyday empirical facts about the causal roles played by certain things in the environment, strongly suggests that some human mental states are externally located. If successful research programs in cognitive science reinforce the judgments yielded by such argumentation, this helps to show that the folk have latched onto genuine natural kinds with their use of 'belief', 'desire', etc. Folk descriptions of these kinds may be more coarse-grained than descriptions of typical cognitive kinds, so the folk kinds may not be cognitive kinds, but perhaps they populate a

---

on 53rd Street. The first explanation supports an extension-friendly ontological reading of the cognitive process in question; the second supports an embedded friendly (and thereby nonextended – see Rupert 2004) ontological reading, according to which all of the cognition takes place inside the organism. I take it that, when faced with such a situation, we should advert to our fine-grained models, given that conditions (a) and (c) are also met. And so far as I can tell, our more articulated (i.e., fine-grained) models virtually all make the distinction I describe below, the distinction between activities of the components of the persisting architecture and the other causal contributors to the production of intelligent behavior, which tends to favor a nonextended view (at least for most subjects, most of the time – see Rupert 2009, 2010).

higher-level domain that bears a systematic relation (say, of supervenience) to the cognitive, and this explains why consilience at the cognitive level helps to confirm folk psychology. One might here add a stronger claim on behalf of the folk categories: they should direct cognitive scientific enquiry, shaping its investigations, and, in some sense, constraining how it is allowed to develop or which theories are to be taken seriously by cognitive scientists.

C. The thought-experiment involving Otto provides an argument for extended mind that has nothing to do with cognitive science or a naturalistic outlook. On this view, any use of the language of natural kinds – at least as I understand them (see note 2, below) – is misplaced. Instead, the nature of mental states is revealed by introspection and conceptual analysis. Cognitive science – with its contrived data and complex, analytical modeling tools – investigates a wholly different domain, one independent of folk psychology, although, for rhetorical purposes, it might be useful to offer arguments for the extended mind and for extended cognition side by side.

Clark and Chalmers seem fairly clearly to have B. in mind. Clark has explicitly endorsed the commonsense functionalist interpretation of the original discussion of Otto and, more generally, a commonsense functionalism that delivers coarse-grained (i.e., very broad and nondetailed) descriptions of the causal roles of mental states that might then be fleshed out in their particular instances by cognitive science (Clark 2008, 88, 96; 2010a; see also Chalmers 2008).

Two comments: First, I agree (Rupert 2004, 422–423) with Chalmers's observation (2008, xii) that certain aspects of folk conceptions of mental states are likely to be at odds with the extended view, and so Clark's methodology should be clarified and defended more pointedly. Second, I think Clark should be dissatisfied with the reason he gives for preferring commonsense to empirical functionalism. He claims that empirical functionalism (or psychofunctionalism) robs us of the possibility of multiple realizations of the mental states (2008, 88 n6), but I can't see how (cf. Wheeler 2010, 260). A system's use of, for instance, the distance between two receptors as a way of computing the location of the source of a sound is a fairly fine-grained functional property (a) that is not part of folk theory, (b) the importance of which in cognition is empirically discovered, and (c) that can be multiply realized – in say, the distance between human ears or the distance between a robot's sound receptors. This example represents the tip of the iceberg: almost every role in any going computational model of actual human cognition is multiply realizable and is the subject matter of an empirical enterprise not constrained in any significant way by folk-psychological commitments. And although most of the states in which cognitive scientists take interest are more fine-grained than belief and desire as Clark conceives of them, we should leave open the possibility that empirical science will find such coarse-grained states as belief and desire to be useful. Such states, being coarse-grained, are particularly susceptible to multiple realizations; more importantly, Clark should remain open to the possibility that, as part of that empirical functionalism, descriptions of the roles of such states might be revised on the basis of empirical results and scientific theorizing (Rupert 2004, 423).

In what follows, I'll focus on A. I find this version of the natural-kinds argument to be of special interest, not only because of its potential connection to the philosophical foundations of cognitive science, but because I'm inclined to think that the methods of the natural sciences, broadly construed, are more likely than are other methods to yield knowledge about mind, self, and cognition (assuming these things exist!). If we would like to know, for example, whether mental or cognitive states extend beyond the boundary of the human organism, we should ask where things stand with regard to our best or most promising scientific theories of mind and cognition. Thus, in my view, some version of the argument from natural (or causal-explanatory)[2] kinds offers more promise than any other argument that's been given in support of the extended view.

---

[2] Henceforth, I omit the qualification 'causal-explanatory'. I include it here partly to indicate how thin a notion of natural kinds can be appealed to in getting Clark and Chalmers's natural-kinds argument off the ground. I assume that for the purpose of understanding the natural-kinds argument, being a natural kind has nothing necessarily to do with average persons' categorization of items they encounter in their natural environment and that it needn't imply the existence of microstructural essences (how could it require the latter? if it were to, the kinds and properties of fundamental physics wouldn't be natural kinds!) or be associated with homeostatic property clusters (Boyd 1991). Natural kinds are simply the causal-explanatory properties and kinds of the successful sciences, or to be a bit more careful, the properties and kinds that our sciences attempt to identify. As such, they are the kinds or properties that ground successful induction (Quine 1969a), appear as *relata* in laws of nature (Fodor 1974), or play causal-explanatory roles (Kitcher 1984). A positivist might insist on a linguistic formulation,

As a final preliminary point, it might be worth remarking on the connection between Clark and Chalmers's thought-experiment involving Otto and the topic of memory. Otto's case involves (dispositional) belief, and thus my use of it to frame questions about memory and cognition might raise eyebrows. Bear in mind, though, that Clark and Chalmers describe Otto's notebook as playing the role of "biological memory" (1998, 12) and throughout their paper repeatedly treat Otto's nonoccurrent belief as a kind of memory or as analogous to it (13, 15–16; and see Clark 2008, 76). This places the discussion squarely in the realm of cognitive science, so long as we take Clark and Chalmers to be talking about memory as it – or what we're inclined to categorize as memory-related behavior – has long been an object of scientific study. So, it does not stretch connections too far to move from natural-kinds reasoning, as applied to the

---

claiming that talk about natural kinds is merely a way of talking about terms that play certain roles in scientific discourse, for example, general terms appearing in covering-law-based scientific explanations. I will not wade further into any of this. Interpreted as philosophers of cognitive science, I take Clark and Chalmers to be suggesting that our best cognitive science will deal in such terms as 'belief' and 'memory' and apply them to states that are at least partly constituted by physical matter beyond the boundary of the human organism, independent of any particular theoretical orientation in philosophy of science. (Because I see natural kinds as something one might roughly label 'scientific kinds', I group what Walter and Kästner [2012] call 'natural kinds' with all other kinds they treat as scientifically legitimate – including certain cluster-based kinds or family-resemblance kinds.)

hypothetical case of Otto, to a discussion of memory, then to cognitive science's search for natural kinds.

II. Natural kinds meet the Parity Principle

In previous work, I pursued this memory-based tack, evaluating the natural-kinds argument by asking whether the science of memory supports a specific instance of the argument; I concluded that, at least so far as one can generalize from this test case, the natural-kinds argument for cognitive extension fails, and does so in an instructive way (Rupert 2004, 405–424).[3] The argument was widely criticized by defenders of extended cognition – partly, I think, because it was widely misunderstood. Below I offer what I hope is useful diagnosis and elaboration but, first, a more careful presentation of the natural-kinds argument and my response to it:

The Natural-Kinds Argument for Cognitive Extension

Premise 1. If the most explanatorily powerful (known) framework for theorizing in a given domain presupposes a given taxonomy of kinds of states, we should at least tentatively accept the existence of states of the kinds in question.

Premise 2. The most explanatorily powerful (known) framework for theorizing about intelligent behavior presupposes kinds that, in fact, have a significant number of instances external to the human organism.[4]

---

[3] Many others have written about memory as it relates to the hypotheses of extended mind and extended cognition: Rowlands 1999, Adams and Aizawa 2001, Sutton 2004, 2010.

[4] One might exclude reference to humans here, but that would be to enter into a different debate. So far as I am concerned, the question is not whether there could be organisms to

Conclusion. Therefore, we should at least tentatively accept the extended view of human cognition.

In response, I argued that premise 2 of the Natural-Kinds Argument falls to a dilemma: either the proponent of cognitive extension individuates the relevant causal-explanatory kinds in a fine-grained way or in a coarse-grained (or generic) way. With respect to the first horn, I argued that we shouldn't expect repositories for external memories to exhibit the fine-grained properties and dynamics (e.g., conversational dynamics) of interest to working cognitive scientists. *Ergo*, our most powerful framework for explaining intelligent behavior in humans – at least if we limit our focus to fine-grained properties and *explananda* – does not seem to presuppose states that have both biologically internal and biologically external instances. My discussion of the second horn was less extensive, but the essential worry was this: characterizations of generic kinds of the germane sort (ones likely to have external instances) would be so thin as to rob them of causal-

---

whom the extended view applies. In my view, the answer to that question is pretty clearly "yes." Neither is the question at hand the question whether there are, in fact, cases of cognitive extension among known nonhuman thinkers. This is a more interesting question, but not at the heart of the debate (some authors appeal to work in robotics when debating extended cognition, partly because such cases are interesting cases in their own right, but primarily because such cases are supposed to shed light on the nature of cognition in a way that has implications for the human cases – see, for example, Wheeler 2005). Thus, in what follows, even when I consider the most far-fetched thought experiments, I treat them as ways of trying to figure out whether humans have extended cognitive states.

explanatory power (one side-effect here being problems of cognitive bloat) (Rupert 2004, sections V-VIII, and for the two horns explicitly presented side-by-side, see *ibid*. 407, 418-19, 424).

The argument drew many responses; some misunderstood the dialectical purpose of the argument, taking it to be a direct attack on the extended view, while others missed the dilemma-structure of the argument, thinking I was somehow on about fine-grained differences only – that I took fine-grained similarity of inner and outer to be necessary either to the extended view or to the natural-kinds argument – while others thought I was on about the Parity Principle (rather than the Natural-Kinds Argument) (Clark 2008, 112–115; Menary 2006, 331, 333, 334, 339–340; Rowlands 2009, 3; Bartlett 2008**,** 171; Sprevak 2009, 506; Levy 2007, 58–59; Adams and Aizawa 2008, 13). Much of the remainder of the present essay is an attempt to straighten out the relation between the Natural-Kinds Argument and the status of memory and cognition as genuine scientific kinds – whether fine-grained or coarse-grained – and how all of this relates to the supposedly extension-supporting role of the Parity Principle.

Early in their paper, Clark and Chalmers assert, "If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process" (1998, 8). This came to be known as the Parity Principle, and on one way of interpreting it, it simply asserts that if something's cognitive, then it's cognitive, regardless of where it sits; as such, it is a truism. If it has any value, it is as a warning against biochauvinistic prejudice.

I agree with Wheeler (2011, 419–420) that this bare, anti-prejudice reading sheds no light on the nature of cognition or on the boundary between what is cognitive and what is not (Rupert 2009, 30–35; Walter and Kästner 2012); instead, it invites the formulation of a location-independent theoretical account of cognition (for an account in this vicinity, see Rupert 2009, chapter 3; 2010). The Parity Principle itself is of no direct use in this process. Beyond the truism that something is what it is regardless of where it's located (provided the change in location doesn't change what it is – Rupert 2009, 33), the Parity Principle seems to ask us to consult pretheoretical intuitions in order to identify a natural, or scientific, kind (*ibid.* 32; 2010, 345–346). Such intuitions may be indispensable when as a means of identifying the *explananda* of cognitive science, but they have little place in the construction and interpretation of theoretical accounts of the *explananda*, and it is here where the natural kind *cognition* (be there such a kind) has its home; cognition is supposed to be the process that produces intelligent behavior – not the behavior to be explained – and the nature of that process should be discovered in the standard way, by the causal-explanatory theorizing and experimentation distinctive of the empirical sciences.

There is another way to gloss the Parity Principle, which might be called a 'benchmarked' interpretation (Wheeler 2010, 254; 2011, 418).[5] On this way of

[5] Wheeler uses the terminology of 'benchmarking' more broadly than I do. He counts a theoretical specification of the nature of cognition as a benchmark (2011, 425), whereas I apply 'benchmark' only when our measure of what is cognitive refers explicitly to the actual states or processes of some specified population, and we thus treat other things as cognitive only when they exhibit the same causal profiles as the relevant states of our

understanding and applying the Parity Principle, one locates an uncontroversial case of a cognitive state or process, then commits oneself to fairness: any place where *that* kind of state or process appears, it is cognitive, even if it's external to an organism. This jibes well with Clark and Chalmers's discussion of Otto (as Clark acknowledges – 2011, 451 – even though he in fact prefers the nonbenchmarked reading of the Parity Principle).[6] They compare our man Otto to a counterpart, Inga, who has intact biological memories; Otto's notebook-based memories play the same coarse-grained roles in the production of

---

reference population. (It complicates matters to include multiple, but not all possible, species in one's benchmarking category, but the essential points remain the same.)

[6] Although Clark prefers the nonbenchmarked interpretation, he appreciates the difficulty of doing without a benchmark (2010a, 54). In the end, though, he thinks we can do without one, by adopting a commonsense functionalism and thinking in terms of the coarse-grained causal roles of everyday mental states (*ibid.*, 55). I'm skeptical, though; philosophical intuitions driven by commonsense functionalism are prone to implicit human benchmarking. There are infinitely more possible coarse-grained roles than there are mental states of the sort we folk associate with our everyday mental lives. So, when folk generate descriptions of coarse-grained roles, they winnow the possibilities somehow, and, most plausibly, they do so by working from their own case (why else care about *these* coarse-grained roles, among the many, many possibilities, except that they're the ones we take to play a role in our own psychologies?). Thus, so far as I can tell, Clark's recommendation that we use such roles to identify the cognitive or the mental does not, in the end, provide a recipe for a nonbenchmarked approach to cognition or mentality.

behavior as those played by Inga's bio-memory, and on this basis – together with the fairness enjoined by the Parity Principle – Clark and Chalmers conclude that Otto has an extended belief.

Thus, the Parity Principle can be applied in a nonbenchmarked or a benchmarked way.[7] The Natural-Kinds Argument might be taken to have two parallel interpretations. Consider a benchmarked reading. Schematically, such reasoning first establishes that certain natural kinds have a significant number of known instances (that is, certain kinds of states, conceived of as having causal profiles of a given, definite sort, play a confirmed role in our best causal explanations of intelligent behavior in the human case); at the next step, it is shown that instances of the same kind (that is, with the same causal profile) appear elsewhere, in locations we might not have expected; and since our causal-explanatory purposes are best served by adverting to these natural kinds in the known cases, the conclusion follows that there are instances of those kinds in locations

[7] Note a further complication in the application of the benchmarked Parity Principle. How do we handle negative cases, in which the benchmark is something we have decided is noncognitive – because, say, if it were in our heads we would count it as noncognitive (Coleman 2011)? Although the Parity Principle is sometimes presented as defense against biochauvinistic prejudice, we apply it at our own peril; for it appeals to, and seems to reify, pretheoretic judgments about the sorts of things that are cognitive – judgments that may well work against the extended view. I think it's best simply to set such judgments aside, for cognitive-scientific purposes (Rupert 2009, 32*ff*; 2010, 345–346), unless, of course, one is engaged in the study of those judgments as behavioral phenomena in their own right.

previously unknown. Within the context of the extended-mind debate, a natural reading of this argument treats internally located human cognitive states as the benchmark – they are the known instances – arguing that our best scientific explanations of human behavior appeal to instances of these kinds characterized in such a way that, surprisingly, they, in fact, have externally located instances. Clark and Chalmers's talk of explanatory unification and more deeply unified kinds suggests this reasoning, as does their comparison of the causal role of Otto's notebook to the causal role of the relevant part of Inga's brain, seeming to use the latter as an argumentative anchor. (This is how I interpreted the natural-kinds reasoning in Rupert [2004], which suggested, to those focused on the Parity Principle, that I was gunning for a benchmarked version of the Parity Principle.)

Perhaps the benchmarked presentation of natural-kinds reasoning is an artifact of Clark and Chalmers's concern about folk mental states; if they are to convince readers that mental states, as the folk understand them, are sometimes externally realized, Clark and Chalmers must convince readers that the external is on par with the internal – given that, for the folk, internal states are paradigmatic mental states. This, however, is a matter of rhetorical strategy (as I think Clark recognizes), not anything to do with the metaphysics of cognition or the correct theoretical gloss of our best cognitive science. Thus, just as Clark has rejected a benchmarked reading of the Parity Principle, we should reject bracket the benchmarked reading of the Natural-Kinds Argument. Recall, though, the costs of Clark's rejection. When we attempt to apply the Parity Principle in a nonbenchmarked way, it becomes clear that how little the Parity Principle has to offer. It says, in effect, "If something plays the cognitive role when it is in one location, then that

thing is cognitive when it plays that role in any other location," without shedding any light on the role the playing of which makes something cognitive. Thus, the move to a nonbenchmarked reading guts the Parity Principle (and in my estimation, this reflects a general shortcoming of the Parity Principle, that it is inert in the debate over extended cognition – Rupert 2009, 31–35).

Does the Natural-Kinds Argument suffer a similar fate, absent benchmarking? Natural-kinds reasoning does not, in the first instance, concern location. It implores us to do our best science of intelligent behavior and *then* check to see whether that science involves kinds that appear in external locations (or, as I would want to describe things, it implores us to do our best science, then check to see whether the results presuppose a deep theoretical difference between two kinds of contributors to the production of intelligent behavior, some of which are (a) privileged over the others and (b) have instances that appear beyond the boundary of the organism). A nonbenchmarked version of Premise #1 of the Natural-Kinds Argument would say something along the following lines: "If our best account of the distinctive drivers of intelligent human behavior invokes natural kinds that have instances beyond the boundary of human organisms, then so be it." On this reading, the argument does not particularly recommend that we attempt to establish the scientific value of kinds with internally located instances, to be used then as reference points in a search for external instances of cognition-related kinds. This version of the Natural-Kinds Argument simply appeals to scientific results, which, at this juncture, strikes me as the most worthwhile sort of natural-kinds reasoning to explore.

Here, then, is an interim summary of the issues and my (sometimes tentative) views about them, specifically in connection with the evaluation of the Natural-Kinds Argument.

Three matters are of primary interest to me here: (1) whether *memory* and *cognition* are natural kinds; (2) whether they, or human-specific versions of them, play a role in the production of intelligent behavior in humans; and (3) whether, if they play a role in the production of intelligent behavior in humans, they have a significant number of instances that appear at least partly beyond the boundary of the human organism. With regard to (1), I'm inclined to think that memory is not a natural kind and that we have little evidence, either way, about the case of cognition. Regarding (2), memory plays no role in the production of human intelligent behavior, but I remain agnostic about cognition; even relativized to human-benchmarked kinds, I don't think generic human memory plays a causal-explanatory role, although human-benchmarked cognition does. Turning to (3), I have taken the question of memory off the table; in the case of cognition, though, our best current explication of human-benchmarked cognition most likely yields little in the way of cognitive extension.

I arrive at these conclusions partly by considering, and rejecting, ways one might pursue alternative, more extension-friendly lines of reasoning. If we focus on fine-grained properties that (a) produce human memory-related behavior, and (b) have been of interest to cognitive scientists (and thus are properties of the contribution of which to the production of human intelligent behavior we have some evidence), it's unlikely that they have extended instances sufficient in number and importance to alter significantly the way cognitive science is done or our fundamental conception of humans as thinkers; this is the upshot of the discussion in Rupert (2004). A proponent of the extended approach might react by attempting to locate organismically external states that fall under more inclusive kinds, say, *memory* or *cognition*. One promising way to prosecute this strategy

argues, first, that internal instances of fairly coarse-grained (or generic) kinds play a role in the production of human intelligent behavior, then, second, to show that some external states satisfy the conditions for membership in those generic kind. In my earlier work, I questioned the value of such generic kinds – expressing, for instance, concern about what has come to be known as cognitive bloat (2004, 421). Much of the remainder of the present essay might be read as an attempt to bolster this skepticism about generic kinds, a matter to which I return in section III, below; but, first, a detour.

III. Methodological interlude: Operationalism, natural kinds, and folk kinds

How do the sciences home in on nature's kinds and properties or select the language to be used in the formulation of their most successful theories?[8] The naturalist's way (Quine 1969b, Quine and Ullian 1970) to answer this question would be to examine what appear to be our most successful scientific theories, checking to see how they were formulated and chosen over competitors. In this section, I focus on the circumstances surrounding the emergence of a new science, when we are faced with what we think is a distinct domain of phenomena the investigation of which demands introduction of new vocabulary, concepts, principles, and laws.

In the present context, the most pressing questions are "what data is the introduction of 'memory' and 'cognition' *as theoretical terms* meant to help us to systematize or explain?" and "why think there are scientific kinds *memory* and *cognition*?" Answering

[8] Hereafter, I omit qualifications meant to emphasize this paper's neutrality in respect of the issue of scientific realism and competing, more procedurally oriented interpretations of scientific enquiry. I ask instead that the reader take this neutrality as read, even when the particular choice of words suggests a thoroughgoing realism.

these questions is no simple matter. Complications arise at every turn – from the identification of data, to the attempt to describe what seems to unite them in one domain of enquiry, to the characterization of the natural kinds supposed to play a substantive role in our best theories of those data.

So far as I can tell, successful scientific (sub)disciplines take shape largely in response to some range of observable phenomena that (a) are unexplained (not currently subject to prediction/retrodiction or manipulation and not currently subsumed under a general theory the principles of which have garnered scientific support from circumstances of successful prediction/retrodiction and manipulation) and (b) seem similar enough such that when they are accounted for, we suspect it will be by some common set of basic principles, theories, or models. What makes us think that given collection of phenomena are of a theoretical piece? It may be partly because we deploy the same everyday terms (e.g., 'remembering' or 'intelligence') in our descriptions of them or, in some cases, because we have had limited success predicting and explaining those phenomena using the same proto-theory or folk theory (for discussion of one relevant case, that of a folk theory of mind, see Gopnik and Wellman 1992). Early thinkers in the field then propose models of the mechanisms that produce these phenomena or hypothesize properties the interactions among instances of which account for the observed phenomena. The target of such modeling is likely, especially in the early days, to be a small subset of the phenomena in the domain. Depending on the degree of success such modeling meets with, a given model, theory, or collection of models might be generalized to account for a wide range of the initial phenomena and, perhaps, after a flash of insight, to account for phenomena that weren't thought to be of the same type as the phenomena to which the

model in question was, initially, successfully applied; additionally, if an otherwise powerful model or family of models does not account for one of the phenomena originally thought to be in the relevant domain, we may expel that phenomenon from the domain. (Here's a toy example: if two chemicals feel to humans to be of different temperatures, but turn out to have such effect because of their chemical interactions with human skin – not because of differences in their mean kinetic energy – we will cease treating the difference between these chemicals as something to be explained by the same sort of models that explain the difference between warm water and cold water). Thus, begins a reciprocal dance: we recategorize phenomena into new similarity-groups based on the similarity of the models (the properties invoked by the models and actual relations claimed to hold between elements in the models) that can successfully account for at least some of them, while the constitution of what seem to us to be current similarity groups (including second-order similarity groups) guides our search for more general models that unify a broader range of phenomena. Along the way, relations of overlap between domains are discovered, single domains split into multiple ones, and the place in the enterprise of everything from individual bits of data to entire phenomena can be reconceived.

The preceding description of the birth and development of science goes all too quickly, running roughshod over the wealth of nuance historians of science have unearthed in recent decades. Nevertheless, this brief description is, I think, accurate as far as it goes, and it makes certain of my commitments clear. Two are of special importance.

The first concerns operationalism: nothing in my sketch of the birth and development of a science entails operationalism of the sort reviled in philosophy of science; in fact,

quite the contrary. According to the operationalist view, a theoretical term expresses nothing more than a claim about the results of measurement-operations that guide the application of that term: if we measure temperature by use of a mercury thermometer, then 'temperature of system x' simply means 'the readout on a mercury thermometer after it's been placed in system x in such-and-such way for n units of time'. Thus, operationalism yields structured sets of observation sentences as the meanings of such terms as 'temperature', 'charge', 'spin', as well as 'remembers'. In order to incorporate distinctively theoretical statements (which undoubtedly play a role in science), these operational meanings might be thought of as inference tickets to move from certain observation sentences ("the thermometer reads n at t") to certain theoretical sentences ("the temperature of the system in question is n at t"), the latter of which can be plugged into formulae that license further claims about the results of measurements ("the thermometer will, at t+10, read n+5").

Although philosophers of science have rejected such orthodox operationalism, its spirit lives on. Many philosophers and scientists remain suspicious of theoretical posits and continue to suspect that terms for theoretical kinds or properties amount to nothing more than tools for categorizing observations or observable phenomena, for instance, that the property of having a memory that P is nothing over and above the subject's exhibiting certain behavior (that is, the behavior we would normally categorize – for practical purposes – as expressing one's memory or acting on one's memory that P). For some, this persisting intuition takes a more liberal form, according to which a term for a given theoretical property or kind applies to a given state if and only if that state is now producing or has before produced the phenomenon of interest.

My sketch of the birth and development of a scientific (sub)domain supports none of these operationalist intuitions, neither the strictly operationalist, nor the more liberal vaguely related, ones. I hope to discover the nature of memory and, more broadly, the nature of cognition, by identifying the privileged theoretical constructs that emerge from the ongoing interplay described above; such an enterprise typically (perhaps always) deals in theoretical constructs that are not understood simply as whatever, in actual cases in which observable phenomena of interest occur, produced those phenomena – and which are *surely* not to be equated in any way with the results of specific measurements. Subjects can have thoughts and memories never voiced or acted upon, and our characterization of what it is to have thoughts and memories should not build in the requirement that every thought or memory actually produce the primary phenomena to be explained. For example, neuroscientific evidence might help to verify the occurrence of a memory-forming process (because we've detected the kind of neural activity hypothesized by our best theory of the production of memory-related phenomena to occur when memories are formed) in a case in which the subject never exhibits any memory-related behavior. (The preceding example is not meant to privilege neuroscience; if the proponents of an extended view of memory are correct, then we might someday have evidence that a given subject has a behaviorally inert memory because we have evidence of the occurrence in her of an at least partly external process that normally occurs when subjects form extended memories.)

The second point is this: the picture of science sketched above leaves plenty of room for a contribution from the folk, even if folk psychological terms do not refer to natural kinds (that is, kinds that will be the subject matter of our mature scientific theories). Such

folk terms as ‘remembers’ and ‘memory’ (and ‘learned’, and ‘meaning’, and ‘concept’ – cf. Machery 2009) may be of use in framing the phenomena originally thought to be of a piece, and thus worthy of investigation as part of the same organized enterprise, even when they ultimately turn out to be bankrupt as kind terms. Moreover, beyond the early grouping together of phenomena, these terms might continue to be used as a matter of convenience – for example, as an easy way of referring to multiple, distinct kinds or to the various products of historically related research projects; in this way, the terms serve something more like a sociological purpose than a purpose internal to the relevant scientific enterprise itself. (Analogy: One might wish to refer to all of the innovations that came out of Bell Labs – perhaps as a way of commenting on the culture at Bell – and thus talk conveniently of the “Bell Labs Research,” without thinking the research thus issued exhibits any deep natural unity.)

IV. Memory, generic memory, and Martian memory

Return, now, to the Natural-Kinds Argument. Premise #2 makes a claim about the taxonomy employed by our best cognitive science and about the location of instances of that taxonomy’s component kinds. Given that natural kinds can, in principle, be either fine-grained or coarse-grained (Rupert 2004, Sprevak 2009, 512), it seems reasonable to try to find a role for coarse-grained, or generic, kinds in our most developed cognitive science – the study of human cognition. In this section, I focus almost exclusively on the possibility that generic memory appears in our best cognitive-scientific taxonomy. But at the end of the section, I revisit questions about more fine-grained properties and kinds.

Memory in humans is a widely disparate phenomenon: different memory systems are likely to be at least as different from each other, in ways that matter to cognitive-

scientific explanation, as they are from other systems not associated with memory-related behavior. It now appears that the best ways to account for (what we might pretheoretically take to be) memory-related *explananda* presuppose a wide range of processes connected in no other way than that each kind of process (a) produces one or more of the behavioral phenomena in question and (b) involves changes in associative strengths of some sort (either described in terms of psychological or neurological mechanisms, not simply in terms of changes in behavior). The presence of the first similarity does not establish the existence of a generic kind; it might suggest that there is such a kind, but the lion's share of argumentative work will revolve around (b). I contend that associative changes are too widespread (they occur in connection with, e.g., lesions and perceptual processing) to ground the inference to a generic kind, *memory* (or even a kind *memory in humans*). In light of cognitive scientists' widespread reference to what they sometimes call various "kinds of memory" (procedural, declarative, semantic, episodic, etc.) and, more importantly, the proliferation of models of various processes operative in connection with these different kinds of memory, skepticism about memory *simpliciter* does not seem unwarranted (Tulving 2000, 41; Michaelian 2010). What considerations might mitigate this skepticism?

Some of Clark's remarks, as well as some of John Sutton's, bear on this issue. On their view, the fact that human memories comprise a motley at the fine-grained level does not preclude the development of a proper science of memory (although note that Clark has since distanced himself from the underlying assumption of a motley: 2011, 452). When discussing the motley of causal processes involved in the production of intelligent behavior, Clark expresses hope for a science of cognition regardless of such disunity:

"The study of mind might...need to embrace a variety of different explanatory paradigms whose point of convergence lies in the production of intelligent behavior" (2008, 95; see also, 2010a, 64).[9] And in advocating for an overarching, interdisciplinary approach to memory, Sutton suggests that we should pursue all manner of memory-related phenomena, looking for "higher-level accounts which do find commonalities" (2010, 214) in spite of disunity at the nitty-gritty levels.

What, then, distinguishes a legitimately scientific generic kind from a merely nominal one (even though the terms referring to such merely nominal kinds may nevertheless serve a useful purpose in scientific discourse). Here is a proposal. As a baseline condition, the various instances of the generic kind must share more than that they are "the kinds of things that produce the phenomena of interest." It will not do, for example, to characterize *generic memories* as "states that participate in processes that produce behavior that matches, along dimensions of content or structure, external material with which the subject has causally interacted" or, in the case of *cognition*, "processing that produce intelligent behavior." These descriptions can, and should, be deployed at the initial stage, as hooks to try to get hold of the natural kinds of *memory* and *cognition*. But, whether there are such kinds depends on whether the hooks do, in fact, get attached to something. More substantively, various instances of natural, generic kinds must bear some kind of family resemblance to each other (cf. Wheeler 2011), but not just any family resemblance: it must be a family resemblance determined by the causal-explanatory roles of the components of the generic kind's instances; there must be a unity

[9] Clark is reacting partly to the specter of eliminativism about the mind or the category of the mental (cf. Sprevak 2009, 522–523).

to various instances that is legitimated by theoretically important aspects of the relevant disciplines themselves. Each instance (or kind of instance) of a generic kind is, I maintain, constituted by a cluster of mechanisms; variations in the components of this cluster, from one species of the generic kind to another, and variations in the relations between these components, determine the fine-grained differences in the causal profiles of various species of the generic kind. Think in terms of partially overlapping models. The models of the way in which various species produce instances of the relevant *explananda* must have significantly overlapping elements and relations among them.[10] This would seem to be the order of the day in most sciences; an initial (typically simple) model of some paradigmatic phenomenon succeeds (well enough), then related phenomena are modeled by the "tweaking" of the initial models – terms are added, parameter values adjusted, etc. If a phenomenon that might have been thought to be of a piece with the others turns out not to be amenable to this "tweak and extend" treatment, it is, and should be, treated as a different kind of phenomenon after all; and this is when we say that original, full range of phenomena weren't all of the same kind – that is, there is no generic kind that subsumes them all.

Now, I agree with Sutton that, generally speaking, we must "wait and see" (*ibid*., 215) what sorts of fruit the interdisciplinary study of memory will yield. But, I take the Natural-Kinds Argument to rest on claims about where we have already arrived. Premise #2 of the Natural-Kinds Argument makes an empirical claim, thereby encouraging us to

[10] It is a question of some interest whether this approach, when applied to memory, yields the same range of natural kinds as Michaelian's tri-level approach to individuation of memory systems (Michaelian 2010, 174).

ask whether our current sciences of memory support that claim. The foregoing proposal gives us some idea what to look for in our search for generic kinds; moreover, I would bet that the motleys in question won't measure up to this standard.

What else is on the table? Wheeler offers the following argument for the scientific legitimacy of generic memory. He asks us to consider the discovery of someone whose inner mechanisms related to "context-sensitive information storage and retrieval" do not exhibit the standard fine-grained human causal profiles. He claims that cognitive psychologists would treat this as "one possible form of the psychological phenomenon of memory" (2010, 258), and that this establishes a genuine role for the kind *generic memory* in cognitive science. According to Wheeler, this hypothetical case establishes the explanatory credentials of *generic memory*, by showing that *generic memory* does "important work in organizing and shaping the project of cognitive-scientific explanation" (2010, 258). I'm not convinced that Wheeler's prediction is correct. But even if scientists would call this subject's processing 'memory', that is neutral with regard to the question of scientific kinds. It is one of the primary theses of this essay that a term's doing such organizational work does not itself show that the term refers to a genuine scientific kind. Moreover, the case is underspecified almost to the point of being a red herring; it may have some effect as an intuition pump, but it's important that we also find an actual property that (a) the average human instantiates, (b) would be doing causal-explanatory work in the case of our hypothetical deviant subject, and (c) does actual causal-explanatory work in the standard human's case.

Consider now the so-called Martian intuition. Here is Mark Sprevak's presentation of the Martian intuition as it pertains to the current debate. "[Rupert's and Adams and

Aizawa's] objection to HEC [the hypothesis of extended cognition] is that fine-grained features of human cognition are necessary for mentality. But this seems wrong. Martians could differ from us in all kinds of fine-grained psychological ways and still have mental states. Therefore, such features are not necessary for mentality" (2009, 509). Although this doesn't quite get my objection right (see above), it does take us quickly to the heart of the matter. If something is a memory if and only if it instantiates a fine-grained kind of the sort found in humans and that participates in the production of memory-related behavior, then Martians don't have memories. Given that our Martians would seem to have memories, we should not use fine-grained human processing as the measure of memory.

Ultimately, Sprevak himself draws a different conclusion, detailed discussion of which here would take us too far afield. I meditate instead on the moral Clark has been inclined to draw from the Martian intuition: it shows (a) that there is a kind of state or process we share with Martians – *memory simpliciter* – the nature of which is coarse-grained enough to subsume the two cases, human and Martian, (b) that *memory simpliciter* is thus a natural kind relevant to the cognitive-scientific explanation of human behavior, and (c) that such a kind is coarse-grained enough that it is likely to have a significant number of instances external to the human organism.

I am not inclined to reject the Martian intuition out of hand. So far as I can tell, though, it is not preposterous to deny memories to our hypothetical Martians (and explain away the intuition). In fact, this seems quite plausible. Assume that Martians exhibit the same kind of behavior we do but produced by mechanisms quite unlike ours. What theoretical grounds might there be for claiming that Martians have memories? Given the

specification of the case, the fine-grained perspective, benchmarked to humans, represents a dead end. One might instead argue that generic *memory* has a causal-explanatory role to play – either (1) in the human case alone (in which case, Martians might have memories simply because they have coarse-grained states similar enough to human generic memories) or (2) in the combined case, by showing that *generic memory* plays a causal-explanatory role when the human and Martian cases are pooled together, even though it doesn't play such a role in the case of humans alone.[11]

We can not rule out (1) and (2) entirely, but it seems to me that a non-kind-supporting explanation of the Martian intuition provides a much more convincing explanation of it than does (1) or (2).[12] On my view, we have the intuition that Martians have memories,

---

[11] Or, one might argue that Martians exhibit a kind of coarse-grained *memory* not found in humans: perhaps there's high-level unity to Martians' various memory systems and a high-level unity among those systems' unity and the unity found among the memory subsystems of still further aliens. This, however, strikes me as pure speculation with no bearing on the human case.

[12] The present paper grows out of a presentation made at the 5th International Conference on Memory, an enormous conference held at the University of York in the summer of 2011. So far as I can tell, although 'memory' appeared in the name of the conference, none of the hundreds of cognitive scientists in attendance reported research on just plain memory, that is, generic memory; this is at least anecdotal evidence that, while 'memory' might play an organizational role, it is not treated as a natural kind of interest in cognitive science. Keep in mind, too, the objections of cognitive bloat to which the notion of *generic memory* is liable to give rise (Rupert 2004, 421).

but only because we can imagine that it would be useful to use the term ‘memory’, under at least some circumstances, were we to come across such Martians and study them scientifically, and it would be useful if for no other reason than the similarity between their behavior and ours. The contrasting approach defies credulity: on that view, even though our models of human-memory-related data don’t make use of a generic kind *memory* – and don’t naturally suggest any such kind when we stand back and consider the similarities between various such models – we should nonetheless force that interpretation on our models because of a pretheoretical reaction to the description of imaginary beings. Whether a given property plays an actual role in our most successful models of human behavior is determined by the models themselves, not by intuitions about hypothetical beings. The intuition that the Martians would have something in common with humans should be treated in the same way I suggested that we treat Wheeler’s intuitions. If humans and Martians exhibit the same forms of behavior (but note how incredibly implausible this is, particularly if one thinks of this as benchmarked to the human behavior normally of interest to memory scientists), there may well be organizational or pragmatic reasons for continued use of the word ‘memory’ (see the “Bell Labs” example above), but this in no way implies the existence of anything with significant ontological (or even methodological) import. We have the intuition that the two species must share something, but most plausibly this is driven by behavioral benchmarking only, and that does not a science make.

I propose, then, that we wait till we find some Martians, and let the chips fall where they may. In particular, let us wait till we discover Martians, then let my theory of generic kinds – itself generated by attention to actual scientific modeling – place a

genuine constraint on the collections of instances (or kinds of instances) that can be of the same generic natural kind. If, for example, a Martian exhibits memory-related behavior, but that behavior is produced by a collection of *very* different mechanisms from the ones that produce memory-related behavior in humans (and the Martian process is not amenable to tweak-and-extend modeling), then the Martian behavior is not produced by memories, at least not if we want to use 'memory' as a natural-kind term, rather than, say, as a merely organizational term.

Return to the question of nongeneric kinds. Might there be an argument from, say, mid-grained kinds in support of the extended view? One of the main thrusts of Rupert (2004) was that benchmarking to fine-grained human states undermines Premise #2 of the Natural-Kinds Argument. There I developed a dilemma that I took to be fatal to the benchmarked version of the Natural-Kinds Argument (*ibid.* 407, 418-19, 424), and in developing the first horn of that dilemma, I argued that the fine-grained kinds of interest to cognitive scientists of human memory are not kinds that we have found or should expect to find in the human environment.

How would the development of that first horn proceed if we were to focus on what might be called 'mid-grained' categories, such as *declarative memory*. Here is how I see such kinds. Although each one helps to explain a relatively large range of phenomena, it does so in a fairly fine-grained way, in that its causal profile is fairly detailed. If we take declarative memory to be the sort of thing variations in the quantity of which or in the characteristics of which account for variance in the relevant behavioral data, then declarative memory turns out to be a somewhat demanding kind; so, benchmarking to humans isn't likely to yield extended declarative memories. Discussion of whether

declarative memory, as a less demanding kind, is legitimate – that is, plays a causal-explanatory role as a scientific kind, not simply as something the nature of which is to produce a certain circumscribed range of phenomena – returns us to many of the issues raised in connection with *generic memory*. How would we show that a kind plays a genuine causal-explanatory role if doesn't correspond to, say, a quantity that appears in a range of successful models? By consideration of alien cases? To take that approach seems to me to strain philosophers' credibility as contributors to the cognitive-scientific enterprise. Such thought experiments might provide interesting psychological data (pertaining to intuitions about thought experiments), but they don't reveal anything about natural kinds; our responses to them don't tell us which properties play a causal-explanatory role in our best existing or emerging cognitive science.

I think that, more fruitfully, we should turn to questions about cognition itself. As emphasized in the preceding section, we can allow that an instance of a given kind, say, *memory*, might never produce the phenomena that kind is hypothesized in order to explain. Nevertheless, it is reasonable to demand more of a genuine memory than that it be the sort of thing that could, under the right circumstances, play a certain role in the production of the phenomena of interest. It might, for example, be reasonable to require that for something to be a memory, it must be cognitive – in which case, we must inquire into the property *being cognitive*. And in this regard, I say that we make cognitive systems our priority (Rupert 2004, 424–428). If we cannot tell exactly what causal-explanatory role, for example, *declarative memory neat* plays in our best models, at least we can ask whether what we think might be a declarative memory appears within the boundaries of a cognitive system (whether extended or not). If it doesn't, this fact counts

against the state's being any kind of memory, given that we're inclined to treat declarative memory as a kind of cognition.

V. *The systems-based account of cognition*

Finally, what about cognition? Is it a natural, or scientific, kind? And if so, how broadly is it instantiated?

As a first step, we should try to understand the nature of cognition in the most salient case, the human one. In other work (2004, 2009, 2010), I have argued (a) that many research programs that treat organisms as containing cognitive systems have been very successful, and (b) that a striking fact about cognitive modeling provides both the best explanation of (a) and grounds a theoretically based location-neutral account of cognition. The striking fact is that virtually all forms of cognitive modeling distinguish between, on the one hand, the persisting architecture, which is taken to have a relatively fixed number of elements (e.g., connectionist units) and stable relations among them (e.g., degrees of inhibition or ways in which the degrees of inhibition change over time), and on the other, more transient causal contributors that, together with aspects of the persisting architecture, produce intelligent behavior (cf. Wilson 2002). Think of this as an inference to the best (available) explanation, twice over. First, the fact that the distinctive and central aspect of cognitive modeling – the persisting architecture – is typically instantiated within the organism explains why there's been as much success as there has been in doing "organism-oriented" cognitive science. Second, that the persisting architecture is the distinctively cognitive thing best explains why it appears in all different forms of modeling.

I do not defend this line of reasoning here but confine myself to some relevant observations about the approach and result it yields:

First, I take it that, for most subjects, at most times, the persisting system is housed within the organism. Note that Otto and his notebook can be described so that, by hypothesis, the notebook becomes part of his cognitive architecture, but that's irrelevant to the claim that we are now undergoing, or have undergone, or should undergo, a revolution in cognitive science; Otto's case is make-believe, after all.

Second, my strategy is to find a common thread among successful model-types, issuing from the entire range of orientations in cognitive science (including modeling that has sometimes been interpreted as extension-friendly): connectionist, computationalist, dynamicist, brute biological, robotics-based, and artificial-agent-based.[13] If this "supervaluation" strategy is right, it provides powerful evidence in favor of the generally nonextended view (so long as the first observation is correct as well). Moreover, this evidence is of precisely the kind that Ross and Ladyman (2010) rightly demand. The evidence has nothing to do with intuitions about causality and constitution or everyday examples involving air conditioning or stereo systems. My claim rests on the real, scientific modeling, on the ground. Note, too, that contrary to what Ross and Ladyman suggest, the supervaluation strategy works: when one considers the full range of successful models, one does not find such a widespread, context-sensitive shift, where

[13] See the last argument in footnote 1 for a complication and, in response, an argument that this complication does nothing to weaken my case for the systems-based view.

sometimes the privileged elements (the architectural elements) are inside the organism and sometimes out.[14]

---

[14] In this context, it might be worth revisiting a worry about such entities as the sun; on the measure I've proposed to diagnose the scope of the persisting integrated architecture (2009, 42–43), something that consistently contributes, along with other mechanisms, to the production of a wide range of forms of intelligent behavior is almost certain to qualify as part of the persisting cognitive system. Objection: doesn't the sun fit that category? Doesn't the big bang? Yes, and although I've tried to avoid these consequences in various plausible ways, it may be best to rely on the "common element" strategy. All forms of models leave, for instance, the sun out of the architecture, and this gives us reason to toss out (by reflective equilibrium, one might say) some mechanisms that might otherwise seem to contribute causally in the way deemed adequate by my formal measure. The models all treat the sun as a background condition and that alone justifies treating it as such.

It might also be worth pointing out that my measure of the clustering of mechanisms (that is, of the scope of the integrated architecture) is consistent with a modular architecture (*contra* the suggestion made by Clark – 2011, 456 – and others). The measure is sensitive to way in which factors co-contribute to the causal production of intelligent behavior, not to whether the factors causally interact with each other when producing that behavior. (One provocative way to think about the mechanisms in question is as local neural mechanisms that perform simple computational functions and contribute to the production of various forms of intelligent behavior partly by being

Third, my approach is locationally neutral. I attempt to find a central and pervasive distinction present in wide range of successful cognitive models, then check, afterwards, to see where the elements on either side of the divide appear. The division is not sought with malice aforethought (my initial reaction to the extended view was actually sympathetic, even though I distinguished it from the embedded view – see Rupert 2001, 505, n7); in fact, the distinction that seems to be most central to cognitive modeling leaves open the possibility that elements on either side of the division fall on either side of the organismic boundary.[15]

Fourth, the present exercise does not presuppose that a mark of the cognitive is needed to do cognitive science. Successful work in cognitive science helps us to see what the mark of the cognitive might be (or what might be necessary conditions for something's being cognitive), but we absolutely do not need a mark of the cognitive to do cognitive science. The present issue is how best to interpret the cognitive science we have (and

---

assembled in different combinations for different purposes (see Anderson 2010), which might or might not be done in a functionally modular way.)

[15] Perhaps the standard *explananda* of cognitive science somehow bias the development of models, so that no matter what orientation one works from, the models developed are more likely than they should objectively be to contain (a) elements of an architectural sort and (b) architectural structures that are more likely than they should objectively be to appear within the boundaries of organisms. Perhaps, but I'd like to see this sort of concern worked out in some detail and accompanied by a plausible alternative suggestion about the *explananda* of cognitive science that would not have this (or any other) biasing effect.

correspondingly, how to interpret claims that cognition extends beyond the boundary of the organism).

Fifth, my proposal fulfills Mark Rowlands's demand for an owner of cognitive states (Rowlands 2009). Rowlands realizes that his various conditions for extended cognition lead to unacceptable results (cognitive bloat, in particular) absent a further constraint on cognitive extension; to be part of a cognitive process it has to be owned by a subject. Naturally, philosophers of cognitive science will want a theory of the self and of ownership. I have offered an empirically motivated one: the self is the cognitive architecture, and it owns a state just in case that state is a state of one of the architecture's component mechanisms.

Sixth, contrary to some of Mark Sprevak's claims, we do not need to settle this issue by appealing to intuitions about hypothetical cases. Sprevak (2010, 361) assigns a substantive role to the Martian intuition, and to intuitions about thought experiments more generally; we must appeal to them if, for example, we are to decide between an extended interpretation of successful empirical work and alternative, less radical interpretations of that work (such as the hypothesis of embedded cognition – Rupert 2004). I disagree. We need a theoretically motivated, location-independent account of cognition to do the job, and we can get that from successful cognitive scientific practice itself. Cognition is whatever distinctive property produces intelligent behavior. I maintain that the property in question is *being the activity of mechanisms that are part of the persisting architecture*.[16]

[16] In a pair of recent papers, Adams and Aizawa (2010) and Clark (2010b) debate the status of a pencil used by a mathematician to solve problems. Adams and Aizawa don't

take very seriously the idea that the pencil itself is cognitive, and, in response, Clark argues that they're asking the wrong question; they shouldn't ask whether the stand-alone pencil is cognitive, as if it's a property the pencil might have in isolation. I think Clark is right, but his being right about this significantly constrains the application of the Parity Principle. The fact that a state or object might be part of a Martian's cognition has no bearing on whether it's cognitive in the human case. Whether or not it's part of human cognition depends on how the human is using it. The state doesn't acquire the property of being cognitive, neat, simply because it's cognitive in a different context, when a Martian interacts with it. In order that the state be cognitive in the case when it's part of a human-centered system, it must satisfy a location-independent criterion for something's being cognitive (or being "human-cognitive," if cognition is not a natural kind) that is sensitive to the state's status on the particular occasion in question. And, on my view, that involves reference to the persisting set of mechanisms that co-contribute, in various overlapping subsets, to the production of a wide range of forms of intelligent behavior.

Although limitations of space prevent a full discussion of complementarity-based (or so called second-wave – Sutton 2010) arguments for cognitive extension, my response invokes the systems-based criterion. It is true that inner and outer contributors to the production of intelligent behavior complement each other in deep ways, but this doesn't change matters with regard to the arguments for the systems-based view. The inner and outer play significantly different roles in cognitive modeling, and that difference is *the* central distinction between different kinds of causes to the production of intelligent behavior. Thus, if there is an interesting distinction between some causes and others, it is to be found on one side but not the other side of the line I have identified.

What, then, of the kind *cognition*? According to the systems-based view, a state (or process) is cognitive (if and?) only if it is the state of a (non-background) mechanism (or is a process made up wholly of causally connected states of various such mechanisms) that is a component of a persisting architecture – that is, a member of the relatively persisting set of mechanisms that co-contribute (although not necessarily by interacting), in various intersecting subsets, to the production of a variety of forms of intelligent behavior (that are part of a single biography). Might this view ground a generic conception of cognition, of the sort suggested in section II? There are, on this view, component mechanisms (a) individual ones of which make systematic contributions to the instantiation of a variety of kinds of *cognition* and (b) manipulation of which can cause systematic variation in what we might recognize as the degree of intelligence of the associated behavior (because what we're *really* out to explain is not just intelligent behavior, but also variations in the degree of intelligence that a given form of behavior exhibits).

Nevertheless, *this is only one kind of cognition*. We simply don't know about other species or other kinds of cognition. Moreover, variations in kinds of cognition, if they exist, will likely involve second-order variation – in kinds of systemic integration perhaps – and may be difficult to get our heads around at present. If there is to be a generic kind, *cognition*, different species (or different sub-groups within species) must produce intelligent behavior via different sets of mechanism that share some highly abstract properties, but not others (properties of dynamics, architectural organization, etc.) that meet the conditions given in section IV for various processes', states', or systems' being different sorts of a generic kind; they must all be such that we can model them by tweak-

and-extend (or tweak-and-simplify) in relation to each other – that is, by accounts that are systematic extensions and variations on my systems-based view. This is a tall order, and the tallness of it provides some reason for skepticism about the existence of generic cognition. To the extent that current cognitive science explores generic cognition, it is in the limited sense in which cognitive scientists attempt to model behavior in nonhuman animals that differs significantly from human behavior. So far as I can tell, such investigations have not produced models that yield different kinds of cognition from human cognition that also exhibit second-order similarities to the human case sufficient to establish an overarching kind *cognition* which itself has instances that are both external to the human organism and play a causal-explanatory role in the production of human behavior.

VI. Afterword

Clark and Chalmers's central claim concerns human cognition and mental states; that's what's so striking about it (we're not where we thought we were!). Above, I express skepticism about the value of far-fetched hypothetical cases in our search for properties that might support the Natural-Kinds Argument. In a recent paper, Clark responds to my systems-based criterion by outlining just such a fanciful case, describing one Metamorpho (2011, 456–457), a creature who travels the world assembling and disassembling "himself" in certain respects during the completion of cognitive tasks. Clark asks rhetorically whether Metamorpho would cognize?

I answer that it depends. Metamorpho doesn't engage in human-benchmarked cognition. Alternatively, he might engage in genuinely generic cognition; Metamorpho's relevant processes and those in humans manifest a single generic kind, *cognition*. The

details would have to be worked out, either in response to the study of a real Metamorpho or a more complete specification of our hypothetical being. If there is no unity, even at the generic level, then surely Metamorpho's way of producing intelligent behavior is irrelevant with regard to much of our reasoning about human cognition. For example, with regard to the application of the Parity Principle, the fact that a human uses something that would be of an altogether different natural kind from human cognition were it to be used by Metamorpho has no bearing on whether it is an instance of *human cognition* when humans use it. But, even if hypothetical Metamorpho would, were he to exist, produce intelligent behavior via processes that are, at a very abstract level, of a piece with human cognition, we should still wonder at the import of that fact. Until we actually discover what sort of properties Metamorpho's system has, and why some of them correspond to interesting aspects of the way we model intelligent behavior in humans, there's no reason to re-interpret existing models of the production of intelligent behavior. At the very least, the ball is in the court of those who say it should. Find those creatures, show that their behavior is produced by processes that share enough with humans' that the two kinds of processes legitimately fall under the same generic kind.[17]

---

[17] Of relevance here is an issue unexplored in discussion of my own systems-based account of human cognition. It is one thing to argue, as I have, that various successful models of human cognition consistently mark a certain distinction. It is another to show that variations along the relevant dimensions make a causal-explanatory difference. What am I imagining? Here's one possibility: take a given collection of mechanisms that count as a persisting cognitive system (by my lights) and calculate the average number of distinct forms of intelligent behavior that each of those mechanisms contributes to the

Think of matters in this way. Sprevak (2009) takes his argument to be a *reductio* of the sort of functionalism that gives rise to the extended view (partly via the Parity Principle). Instead, I take his argument to be an injunction against only commonsense analytic functionalism. A functionalism driven by scientific enquiry itself uncovers natural kinds only where they actually are, while helping us to correct folk intuitions that are off base. In contrast, analytic functionalism most probably endorses empirically useless generic kinds, and we shouldn't tie ourselves in knots trying to accommodate intuitions concerning the application of the corresponding terms and concepts. I say we walk through the open door of naturalism: admit that either human-benchmarked cognition is the only kind of cognition there is, and so cognitive science is about that actual kind (even though the term 'cognition' might have broader non-kind-revealing pragmatically driven application), or hold that there may be other natural kinds of cognition, and so *cognition* itself may be a legitimate generic kind, but accept that we have no access to them at present and should not attempt to interpret our cognitive science as if it were a science that covers or invokes *those* kinds of processes.

---

production of; then ask whether variations in that quantity, from one cognitive system to another, account for variance in the relevant forms of behavior. Ultimately, I suspect that vindication of human cognition, as a natural kind, requires a positive result of this sort; similarly, we should want this kind of result if we are to accept that the (hypothetical) abstract similarity between human cognition and Metamorpho's cognition marks the causal-explanatory contribution of a more abstract kind, *generic cognition*, to human behavior.

References

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14, 43–64.

Adams, F., & Aizawa, K. (2008). *The bounds of cognition*. Oxford: Blackwell.

Adams, F., & Aizawa, K. (2010). Defending the bounds of cognition. In Menary (2010), 67–80.

Anderson, Michael L., (2010) Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33, 245–313.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence* 48, 57–86.

Bartlett, G. (2008). Whither internalism? How internalists should respond to the extended mind hypothesis. *Metaphilosophy* 39, 2, 163–84.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies* 61, 127–148.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, Vol. RA-2, 14–23.

Chalmers, D. (2008). Foreword to Andy Clark, *Supersizing the Mind*. Oxford: Oxford University Press, *ix–xvi*.

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

Clark, A. (2010a). Memento's revenge: The extended mind, extended. In Menary (2010), 43–66.

Clark, A. (2010b). Coupling, constitution, and the cognitive kind: A reply to Adams and Aizawa. In Menary (2010), 81–99.

Clark, A. (2011). Finding the mind. *Philosophical Studies* 152, 447–461.

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

Coleman, S. (2011). There is no argument that the mind extends. *Journal of Philosophy* 108, 2, 100–108.

Fodor, J. (1974). Special sciences. *Synthese* 28, 77–115.

Giere, R. (2006). The role of agency in distributed cognitive systems. *Philosophy of Science* 73, 710–719.

Gopnik, A., and Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind & Language* 7, 145–171

Kitcher, P. (1984). Species. *Philosophy of Science* 51, 308–33.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence* 73, 31–68.

Kirsh, D., and Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science* 18, 513–49.

Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

Machery, E (2009). *Doing without concepts.* Oxford: Oxford University Press.

Menary, R. (2006). Attacking the bounds of cognition. *Philosophical Psychology* 19, 329–44.

Menary, R. (Ed.) (2010). *The extended mind.* Cambridge, MA: MIT Press.

Michaelian, K. (2010). Is memory a natural kind? *Memory Studies* 4, 2, 170–189.

Quine, W. (1969a). Natural kinds. In *Ontological Relativity and Other Essays* (New York: Columbia), 114–38.

Quine, W. (1969b). Epistemology naturalized. In *Ontological Relativity and Other*

Essays (New York: Columbia).

Quine, W., and Ullian, J. (1970) *The web of belief.* New York: Random House.

Robbins, P., & Aydede, M. (Eds.) (2008). *Cambridge handbook of situated cognition*. Cambridge: Cambridge University Press.

Ross, D., and Ladyman, J. (2010). The alleged coupling-constitution fallacy and the mature sciences. In Menary (2010), 155–166.

Rowlands, M. (1999). *The body in mind: Understanding cognitive processes*. Cambridge, MA: Cambridge University Press.

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

Rowlands, M. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. Cambridge: MIT Press.

Rupert, R. (2001). Coining terms in the language of thought: Innateness, emergence, and the lot of Cummins's argument against the causal theory of mental content. *Journal of Philosophy* 98, 499–530.

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

Rupert, R. (2009). *Cognitive systems and the extended mind*. Oxford: Oxford University Press.

Rupert, R. (2010). Extended cognition and the priority of cognitive systems. *Cognitive Systems Research* 11, 343–356.

Sprevak, M. (2009). Extended cognition and functionalism. *Journal of Philosophy* 106, 503–27.

Sprevak, M. (2010). Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science* 41, 353–362.

Stich, S. (1996). *Deconstructing the mind*. Oxford: Oxford University Press.

Sutton, J. (2004). Representation, reduction, and interdisciplinarity in the sciences of memory. In H. Clapin, P. Staines, and P. Slezak (Eds.), *Representation in mind: New approaches to mental representation* (Elsevier), 187–216.

Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In Menary (2010), 189–225.

Thelen, E., & Smith, L. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Tulving, E. (2000). Concepts of memory. In E. Tulving and F. Craik (Eds.), *The Oxford Handbook of Memory* (Oxford: Oxford University Press), 33–43.

van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy* 92, 7, 345–381.

Walter, S., and Kästner, L. (2012) The where and what of cognition: The untenability of cognitive agnosticism and the limits of the Motley Crew Argument. *Cognitive Systems Research* 13, 12–23.

Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT Press.

Wheeler, M. (2010). In defense of extended functionalism. In Menary (2010), 245–270.

Wheeler, M. (2011). In search of clarity about parity. *Philosophical Studies* 152, 417–425.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review* 9, 625–36.