Biol Philos (2012) 27:545–570
DOI 10.1007/s10539-012-9308-4

How to demarcate the boundaries of cognition

David Michael Kaplan

Received: 1 June 2011 / Accepted: 6 January 2012 / Published online: 19 February 2012

Abstract Advocates of extended cognition argue that the boundaries of cognition span brain, body, and environment. Critics maintain that cognitive processes are confined to a boundary centered on the individual. All participants to this debate require a criterion for distinguishing what is internal to cognition from what is external. Yet none of the available proposals are completely successful. I offer a new account, the mutual manipulability account, according to which cognitive boundaries are determined by relationships of mutual manipulability between the properties and activities of putative components and the overall behavior of the cognitive mechanism in which they figure. Among its main advantages, this criterion is capable of (a) distinguishing components of cognition from causal background conditions and lower-level correlates, and (b) showing how the core hypothesis of extended cognition can serve as a legitimate empirical hypothesis amenable to experimental test and confirmation. Conceiving the debate in these terms transforms the current clash over extended cognition into a substantive empirical debate resolvable on the basis of evidence from cognitive science and neuroscience.

Keywords Extended cognition · Embodied cognition · Mutual manipulability · Intervention · Mechanism

Introduction

One important aspect of the philosophical debate over extended cognition (hereafter, EC) centers on how to demarcate the boundaries of cognition. The boundary demarcation problem at the core of the debate has commonly been

D. M. Kaplan (✉)
Department of Anatomy and Neurobiology, Washington University School of Medicine, Box 8108, 660 South Euclid Avenue, Saint Louis, MO 63110, USA
e-mail: kaplan@eye-hand.wustl.edu

framed by proponents and detractors alike as proprietary to cognitive systems (Adams and Aizawa 2001, 2008; Rowlands 2010). In this paper, I argue that this problem is more fruitfully understood as a specific instantiation of the general problem of demarcating mechanism or system boundaries. I demonstrate how elements from the apparatus of mechanistic explanation (Bechtel 2008; Bechtel and Richardson 1993; Craver 2007a, b; Machamer et al. 2000) both provides an independently motivated criterion for determining cognitive boundaries (and accordingly, whether such boundaries extend beyond the brain into the non-neural body and local environment), and helps to transform the current conflict over EC into an empirical debate resolvable on the basis of evidence from cognitive science and neuroscience. After assessing a number of alternative criteria, I propose an intervention-based criterion according to which mechanism or system boundaries are determined by relationships of mutual manipulability between the properties and activities of putative components and the overall behavior of the mechanism in which they figure. The mutual manipulability criterion outperforms all extant criteria because it alone is (a) capable of distinguishing components of cognition from causal background conditions and lower-level correlates, and (b) showing how the core hypothesis of extended cognition can serve as legitimate empirical hypotheses amenable to experimental test and confirmation.

According to the central hypothesis of EC, some cognitive processes (broadly construed to include perception, memory, decision making, learning, etc.) are in part composed of, or constituted by, bodily and environmental structures and processes. In this specific sense, the boundaries of cognition may be said to extend beyond the brain into parts of the non-neural body and surrounding environment (Clark 2008; Clark and Chalmers 1998; Rowlands 1999, 2010; Wheeler 2010; Wilson 2004). Critics deny this hypothesis and instead maintain that cognition is exclusively a neural phenomenon, occurring only in structures and processes internal to the brain and thus confined to a boundary centered on the individual (Adams and Aizawa 2001, 2008, 2010; Rupert 2004). Importantly, most critics of EC unhesitatingly grant that cognitive processes occurring in the brain may exhibit various kinds of causal dependence on wider bodily and/or environmental structures for their proper functioning. Embracing causal dependence does not, however, require acceptance of the more radical claim that these causally relevant factors are genuine constituents or component parts of cognition. In fact, the hypothesis that cognitive processes causally depends upon neural, bodily, and environmental factors leaves much of the traditional conception of cognition untouched because one can continue to maintain that cognition occurs exclusively in the brain.

Consider a widely discussed case in the EC debate (originally due to Rumelhart et al. 1986; also discussed by Adams and Aizawa 2001, 2008; Clark and Chalmers 1998; Wheeler 2010). Computing the products of large numbers without an electronic calculator typically involves the use of pencil and paper. With these simple environmental tools, we can iteratively apply the partial products method, efficiently keeping track of the intermediate computations along the way until a solution is reached. All parties to the debate agree that normal cognitive performance causally depends on these external resources. The debate ensues about how to interpret their contribution. Proponents of EC argue that there are

conditions under which bodily and environmental resources (e.g., pen, paper, and inscriptions) bear more than mere causal relationships to internal cognitive processes (e.g., computing large products). In particular, they maintain these entities should be counted as genuine components or constituents of cognitive processes, and that in such cases the entire system spanning brain, body, and local environment implements cognition. Critics of EC assert that only the brain serves as the proper implementation base for cognition, the rest of the non-neural body and surrounding environment at best operating as non-cognitive, causal contributors to normal cognitive performance.

In spite of their disagreements, all participants in the debate require a set of conditions or *criterion* for distinguishing what is internal to cognition from what is external. The central challenge is thus to provide an adequate solution to what I will call the *boundary demarcation problem* for cognition. This is the problem of defining the internal-external boundary according to which genuine components of cognitive mechanisms, systems, or processes can be distinguished in a principled way from parts of the brain, non-neural body, and embedding environment that merely contribute as the supportive substrate or causal background conditions against which normal cognitive processing occurs in the brain.[1] I am not alone in casting the pivotal issue in the debate over EC in this way. In a key paper, Haugeland (1995/1998) characterizes the central challenge for defenders of EC as providing a principle for "dividing systems into distinct subsystems along nonarbitrary lines." (211). In doing so, Haugeland similarly locates the problem of delimiting cognitive boundaries in the broader context of delimiting system boundaries.

It is crucial to recognize that the debate over EC is not terminological—that is, about what to *call* bodily and environmental factors relevant to explaining cognition. On the contrary, the theoretical issue at the center of the debate has substantive and far-reaching consequences for cognitive science as a whole. If the hypothesis of EC is correct, this directly calls into question the veracity of one of the most cherished theoretical assumptions in cognitive science and neuroscience—that cognition occurs exclusively within the confines of the brain—which transcends computational, connectionist, and even some dynamical approaches. The EC debate thus really concerns whether facts about physical embodiment and environmental embedding can be assimilated into traditional cognitive science or whether these findings necessitate radical changes to both its subject matter and theoretical framework.

The paper is organized as follows. First, I briefly discuss two prominent demarcation criteria internal to the EC debate to highlight difficulties arising when the debate is carried out in terms of demarcation criteria proprietary to cognition. Second, I locate the problem of determining the boundaries of cognition in the broader context of determining mechanism boundaries generally. Third, I assess a

[1] Throughout the paper, I will use 'component' to refer specifically to parts of a mechanism, system, object, or process bearing appropriate relationships of mutual manipulability to the phenomenon as a whole ("The mutual manipulability criterion" section). Consequently, not every arbitrarily subdivided part of a given object will count as a component. See Craver (2007b) for similar treatment. One exception will be in "The Simon-Haugeland bandwidth criterion" section, where Haugeland's alternative views about individuating system components and boundaries are discussed.

highly plausible, generic criteria for demarcating mechanism boundaries, but show that it too has shortcomings. Fourth, I introduce the mutual manipulability criterion and outline its main advantages. Finally, I demonstrate the utility of the mutual manipulability account by applying it to a number of putative examples of EC. In doing so, I show how this criterion helps to transform the philosophical debate about cognitive extension into a tractable, empirical one.

Proprietary demarcation criteria

The debate over locating the boundaries of cognition has largely ignored broader perspectives for delimiting mechanism boundaries in general. Instead, the debate has proceeded under the assumption that there is a plausible proprietary answer to the boundary demarcation problem exclusive to the subject matter of cognition. To be clear, a proprietary demarcation criterion for cognition specifies a set of conditions that must be met for a process to count as cognitive, whereas a proprietary demarcation criterion for biological processes would specify a set of conditions that must be met for a process to count as biological. A generic demarcation criterion ("Demarcating mechanism boundaries" section), by contrast, specifies conditions for boundary demarcation independently of special assumptions about the nature of the system whose boundaries are being demarcated. The idea of a proprietary, cognition-specific criterion is well captured by Rowlands (2010), who characterizes such a criterion as one that "tells us when a process counts as a cognitive one" (108). The main problem with the predominant criteria on offer within the EC debate is that they rely on specific theoretical assumptions about the nature of cognition that are either ill-defined, and therefore difficult to maintain, or are likely to be viewed by opponents as equally suspect as the claims about cognitive extension they are intended to support.

Consider first a criterion for demarcating cognitive boundaries emerging from the critical literature on EC. Adams and Aizawa (2001, 2008) propose a demarcation criterion based on processes that involve the transformation and manipulation of representational states bearing non-derived contents. They argue that application of this criterion effectively disqualifies putative cases of cognitive extension. The main difficulties facing this criterion stem from its reliance on the ill-defined notion of non-derived content.[2] The intuitive idea is that natural cognitive systems are distinctive in having states with content that is not the product of exogenous assignment (i.e., by some other external system or observer), but instead is intrinsic to the system itself. The problem is that, besides loosely connected intuitions about "original intentionality" (Searle 1980), no satisfactory explication of the notion is in the offing. Proponents, including Adams and Aizawa, have been limited to ostensive definition. As things stand, no consensus exists about the conditions under which

[2] The second aspect of their criterion—that cognition involves a distinctive form of causal process—is also problematic, and has been subject to a number of challenges including Clark (2005, 2008). For brevity, I focus only on problems associated with the first aspect.

ascription of non-derived content over derived content is warranted, and well-known challenges for the notion of original content remain unanswered.[3]

A second problem with defining one's demarcation criterion in terms of non-derived content is that there is no agreement on how such contents might arise in natural cognitive systems in the first place (Shapiro 2009). Granting that a coherent notion of non-derived content is available, the task of developing a naturalistically acceptable account of how semantic content emerges in and can be supported by states of physical systems remains. Without such an account it is difficult if not impossible in principle to determine what kinds of systems have such contents or might be expected to have such contents. Adams and Aizawa offer no new answers here, instead relying on hopeful appeals to a range of naturalized theories of content including informational semantics, asymmetric dependence theory, and isomorphism-based views. This is more bluff than argument, however, since the difficulties with each of these accounts are widely recognized and none hold the status of received theory. In a telling passage, Adams and Aizawa admit that "philosophers and psychologists have yet to develop a theory of naturalized semantics that enjoys much widespread acceptance. It remains unclear just exactly what naturalistic conditions give rise to non-derived content; hence it remains correspondingly unclear just exactly what objects bear non-derived content" (2008, 55). This admission counteracts much of the remaining force behind their proposed criterion. Tying a demarcation criterion for cognition to non-derived content appears to require nothing less than a complete theory of content, something that continues to elude philosophers of mind after several decades of effort. Given its contested status and heavy theoretical baggage, the notion of non-derived content represents a precarious foundation from which to construct a boundary demarcation criterion for cognition.

On the other side of the debate, EC proponents have offered their own proprietary answers to the boundary demarcation problem. The demarcation criterion of functional parity articulated by Clark and Chalmers (1998) is open to similar challenge because its plausibility hinges on prior acceptance of a number of theoretical commitments any one of which might be seen as problematic or question-begging by opponents in the debate. The criterion of functional parity, which Clark and Chalmers have dubbed the *parity principle*, states: "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" (Clark and Chalmers 1998, 4).[4] According to one prominent interpretation, especially common among critics of EC, the principle embodies a similarity-based criterion for when an external structure or process should be counted as cognitive (see Fodor 2009; Rowlands 2010). If the functional equivalence between external components (e.g., parts of the non-neural body or embedding environment) and

[3] See Dennett (1987) for original critical discussion. For more recent discussion of non-derived content in the context of the EC debate, see Clark (2005, 2008). For a rebuttal, see Adams and Aizawa (2008).

[4] It should be acknowledged that Clark (2008) explicitly denies that the parity principle operates as anything more than a heuristic device or "intuitive probe" to loosen commitments to the idea of a skin/skull boundary for cognition. However, it is as close as defenders of EC come to offering an explicit demarcation criterion. Rowlands (2010) is one notable exception.

internal neural components of a cognitive processes (e.g., memory consolidation mechanisms in the hippocampus) can be shown, this should endow the external component the same cognitive status as the internal component.

The primary problem with the parity principle as an adequate demarcation criterion for cognitive boundaries is that it requires a controversial assumption about the appropriate level of grain at which functional parity should be construed and assessed. Rupert (2004), for example, argues that the processes involved in proposed cases of extended memory would differ in fine-grained, but nonetheless crucial, ways from memory processes implemented in the brain, and would therefore fail to qualify as cognitive according to the EC proponent's own criterion. Wheeler (2010) and Clark (2007, 2008) counter by arguing that the defender of the parity principle can concede the existence of these fine-grained functional differences, while at the same time denying that such differences entail a breakdown of parity. The benchmark for parity demanded by the parity principle, they maintain, is not sameness of fine-grained functional profile, but rather functional equivalence at some higher level of abstraction. Nevertheless, an adequate, positive answer concerning the right level of grain at which parity is to be fixed is nowhere offered by Rupert, Clark, or Wheeler. The problem for a demarcation criterion based on functional parity is that it is difficult to see how broad agreement might ever emerge about the right level of grain for characterizing functional roles. There simply appears to be no independent reason for accepting fine- versus coarse-grained functional individuation, and consequently, little hope of establishing an uncontroversial benchmark according to which functional equivalence of external and internal components can be assessed.

The parity principle has also been criticized for its reliance on an assumption of functionalism about mental states and processes (Rupert 2004; Wheeler 2010). Functionalism within the philosophy of mind takes many diverse forms, but the common idea underlying this diversity is that a psychological state or process is of a particular type, not because of how it is physically implemented, but in virtue of the set of functional relations it bears to sensory inputs, other internal states, and behavioral outputs (Putnam 1975). As long as the network of functional relations is preserved, the psychological state (and more generally, the cognitive processes or mechanisms supporting those states) can be implemented or realized in an indefinite number of physical substrates including, but not limited to, the human brain. This core functionalist commitment to multiple realizability is crucial to getting the parity principle off the ground because bare functional equivalence of an external component, independently of how or *where* it is implemented, is sufficient to endow it with the same cognitive status as the internal brain component. Although functionalism in some form or other is widely accepted, it is not universally held (e.g., Polger 2004). It seems that the parity principle depends on a potentially contentious theoretical assumption about functionalism that the critic of EC might reasonably deny. Whether or not any of these objections are ultimately decisive against the parity principle remains to be seen. Nevertheless, it faces serious challenges from multiple directions. For this reason alone it is difficult to endorse as a criterion for demarcating the bounds of cognition.

Rather than try to defend either of these proprietary criteria or dog pile on and reinforce the critical assault against them, I instead want to pursue an alternative strategy by proposing a *generic* mechanistic criterion for demarcating the bounds of cognition. Because the criterion based on mutual manipulability relations I ultimately defend is motivated independently of specific assumptions about cognition, it avoids the shortcomings outlined above. In the next section, I motivate this broad shift in strategy.

Demarcating mechanism boundaries

At its core, the issue of demarcating (and extending) cognitive boundaries is a general one. Researchers across many scientific disciplines, including but not limited to the cognitive sciences, struggle with how to delineate mechanism boundaries. Sometimes scientists advance hypotheses concerning "extended" mechanisms. In doing so, they must contend with the problem of partitioning from among the set of causally relevant factors those that are legitimate components of a given mechanism or system from those that play roles as part of the causal background.

To motivate this shift in perspective, consider a case from biology of an animal from a nearly maximally distant vertebrate taxa from humans—the gecko—and a nearly maximally disparate set of abilities from those germane to human cognition—their climbing abilities. Geckos possess remarkable, gravity-defying climbing abilities that are achieved thanks in large part to microscopic hair-like structures called setae that uniformly panel the toe pads of their feet and make contact with the climbing substrate. With approximately 6.5 million setae covering the body surface, it has been estimated that an individual gecko is capable of generating around 1,300 Newtons of shear force[5]—enough to support the weight of two medium-sized people (Autumn 2006). Although researchers knew that setae must somehow be involved, the specific mechanism of adhesion remained unknown for some time. A wide range of mechanistic hypotheses were tested including glue-like secretions, friction, suction, electrostatic attraction, and micro-mechanical interlocking. The current consensus is that van der Waals forces operating between individual foot-hairs and the environmental climbing substrate underlie the gecko's remarkable abilities. The observation that setae tips are structured to flatten out when pressed against a surface indicated that some form of intermolecular bonding was involved because such flattening greatly increases overall contact area (i.e., the area within which intermolecular bonding can occur). Experimental evidence that setae (which are extremely hydrophobic) can adhere equally well to hydrophobic and hydrophilic surfaces specifically confirmed the van der Waals hypothesis, because no other intermolecular forces (e.g., water-based capillary forces) are capable of inducing bonding between two hydrophobic surfaces. Final support for van der Waals adhesion came from the observation that by increasing the surface

[5] Shear force is defined as a force applied parallel or tangential to the surface of a given material. In this context, it increases the animal's resistance to sliding along a surface.

energy[6] of the environmental substrate the adhesive force of the whole gecko increases proportionally (Autumn et al. 2000).

How, then, should biologists go about demarcating the boundaries of the gecko adhesion mechanism? There is a natural inclination to fix the internal-external mechanism boundary by appealing to spatial or compartmental boundaries such as the organism's skin or some other relevant structural or physical barrier centered on the individual. Using such a criterion is a common tactic for demarcating system or mechanism boundaries generally, both within scientific and ordinary contexts, and has considerable intuitive force. Examples of mechanism boundaries coinciding with spatial boundaries are easily found. Mechanisms of internal combustion occur within engine compartments, toasting mechanisms occur within the walls of toasters, DNA transcription mechanisms occur within the boundaries of nuclear membranes, and human digestion mechanisms are located within body cavities. Although a spatial criterion is sometimes sufficient to determine system boundaries, it is not necessary. For example, the neural mechanism underlying action potential generation fails to respect the relevant spatial boundaries of the neuron. Voltage-sensitive ion channels in the neuronal membrane are crucial mechanism components, allowing the flow of ions in and out of the neuron, yet they perform this function precisely because they span across the intracellular compartment of the neuron.[7]

Despite the intuitive force of a spatial criterion, little else justifies drawing the boundaries of the adhesion mechanism to align with the interface between gecko and environment defined by its skin. More significantly, the fact that individual foot-hairs operate by van der Waals forces directly weighs against this intuitive boundary delineation. Because such forces are extremely weak at anything greater than atomic distances, they require proximate contact between the adjacent adhesive and the substrate surfaces. Consequently, the presence of these microscopic forces underlying gecko adhesion abilities depends not only on properties of the gecko, but also on properties of the environmental substrate, as well as relational aspects such as their relative distance from one another. In light of these considerations, a spatial criterion appears to be difficult to maintain. Indeed, there appears to be some justification for the claim that the gecko adhesion mechanism is spatially distributed to include external components spanning the boundary between the gecko and its environment.

The debate over locating the boundaries of cognition is analogous in two respects. First, it too fundamentally concerns how to demarcate mechanism boundaries and whether there is a plausible criterion by which to do so. The fact that the EC debate focuses on different environmentally-extended mechanisms underlying different capacities is of little consequence. Second, a spatial criterion is similarly ill-suited to resolve debates about cognitive boundaries, since the EC perspective emerges out of an explicit rejection of it.[8]

[6] Surface energy describes the disruption of intermolecular bonds occurring at the surface of a liquid or solid.

[7] See Craver (2007b) for further discussion.

[8] Clark and Chalmers emphatically announce: "[w]e cannot point to the skin/skull boundary as justification [for the boundaries of cognition], since the legitimacy of that boundary is precisely what is at issue" (1998, 8). Critics of EC are thus understandably reluctant to employ spatial criteria to delimit cognitive boundaries so as to avoid charges of question-begging.

The Simon-Haugeland bandwidth criterion

As a potentially viable alternative to the spatial compartment criterion, one might instead define the boundaries of cognition in terms of the pattern of causal interactions among parts within a given system. Herbert Simon (1969) first suggested this as a general criterion applicable to any complex system—artificial or biological. According to Simon, the components and overall boundaries of a system (defined mathematically in terms of state variables and dynamic equations governing how its state evolves over time) can be delimited by taking into account the "intensity of interaction" or causal connectedness between parts in a system. By intensity of interaction, he means the degree to which the behavior of each part in a system is affected by the other parts. Simon argues that boundaries can be understood to fall where the intra-systemic causal interactions (expressed by mathematical dependency relations in the dynamic equations) within one set of state variables are stronger than extra-systemic interactions (i.e., interactions with state variables outside that set). Expanding on this account, Haugeland (1995/1998) defines systems as "relatively independent and self-contained composites of components interacting at interfaces" (1998, 215). Interfaces are in turn defined by Haugeland as points of contact between systems (or subsystems) in which the interactions are "well-defined, reliable, and relatively simple" (1998, 215). Therefore, according to the Simon-Haugeland criterion, an arbitrary part or subsystem counts as a genuine component when the causal interactions within that part are appreciably greater than interactions between it and other parts of the system. Following Grush (2003), we can call this the *bandwidth criterion* to highlight the fact that components within a system (and overall system boundaries) are defined by a particular causal interaction or bandwidth profile—dense or high-bandwidth connections within a given system or component and sparse or low-bandwidth interfaces to other systems or components.

Haugeland's aim in appealing to the bandwidth criterion is to pave the way for EC. Haugeland claims that the traditional decomposition of the dense causal loop into one discernible component—the brain—neatly interfaced between two other components—the body and surrounding environment—cannot be maintained on principled grounds provided by the bandwidth criterion. As a result, cognition extends into the non-neural body and surrounding environment. He offers two interrelated arguments for this claim. First, since system boundaries are, according to Haugeland, defined by the presence of low-bandwidth interfaces, establishing that the causal interactions between brain, body, and local environment are high-bandwidth in character is an argument *against* their being separable components and *for* EC. He thus asserts that "[t]here are tens of millions (or whatever) of neural pathways leading out of my brain (or neocortex, or whatever) into various muscle fibers in my fingers, hands, wrists, arms, shoulders, and so on, and also from various tactile and proprioceptive cells back again" (1998, 225).

Second, Haugeland argues that in virtue of their low-bandwidth interfaces, genuine components should be replaceable in principle with functional equivalents, as long as the overall functional profile is preserved in the newly plugged-in component. As he puts it, low-bandwidth "interfaces provide a natural point of

subdivision—in the sense that any alternative output system with the same behavior could be substituted without making any significant difference" (1998, 224). Crucially, this indicates that the precise implementation details of the components on either side of an interface are relatively unimportant. For example, a resistor in an electric circuit can be replaced with any other resistor (with the same resistance value) because its interaction with the rest of the circuit components are through simple, extremely low-bandwidth connections. Moreover, their functioning does not depend on any properties of the resistor (e.g., its color or mass) except for how much current it resists. Haugeland wishes to deny that this sort of view applies to brain-body interactions. He argues that a complete switching of the output device to which motor signals from the brain are sent—from one's own limb to another human's limb or a prosthetic limb—is not possible in principle. Haugeland instead maintains that structures and processes involved in sensorimotor control fail this test for componency status, and because of the specificity and implementation-dependent nature of their interactions there might be "no way to 'factor out' the respective contributions of these different dependencies" (1998, 225). Haugeland claims that the complex interplay between the relevant patterns of motor activity in the brain and the limb being controlled precludes the possibility that they interact through a simple interface. Instead, they are supposed to exhibit a kind of specificity of interaction tied to the details of one's physical embodiment. These details might include factors such as how the musculoskeletal system imposes biomechanical constraints on degrees of freedom of motion and movement dynamics. For example, our bodies might set limits on hand speed or acceleration because of the maximum torques the system is capable of generating or on how rapidly they can be altered (Nelson 1983). If true, this kind of interdependence makes the prospects for replaceability by functional equivalents extremely improbable.

On the basis of these considerations, Haugeland concludes that there are no real interfaces to align with and justify the traditional boundaries between brain, body, and environment. Because no principled division of the brain-body-environment loop into components interacting at narrow-bandwidth interfaces is feasible, the brain cannot be the component exclusively or distinctively responsible for supporting cognition—since it is not a discernible component in the first place. According to the bandwidth criterion, then, cognition extends.

There are several major problems with the bandwidth criterion as a means to delimit the boundaries of mechanisms in general and cognitive mechanisms in particular.[9] The first general difficulty is that interfaces need not be low-bandwidth forms of causal coupling as Haugeland maintains. Clark (2008) argues that while sparse or low-bandwidth coupling is perhaps sufficient to yield an interface boundary, such a bottleneck is not a general requirement on interfaces.[10] Instead all that is required is some point of contact through which distinct, relatively modular systems can be hooked up and functionally integrated for some period of time and then potentially uncoupled. Clark illustrates this point with the example of so-called

[9] Adams and Aizawa (2008) pose several additional challenges for Haugeland's criterion. Adequately addressing their arguments against Haugeland's account, however, goes beyond the scope of this paper.

[10] Thanks to Georg Theiner for bringing Clark's discussion to my attention.

grid or distributed computing, which involve multiple individual computers networked together to perform large, computationally demanding tasks requiring large numbers of processing cycles (e.g., predicting how proteins fold into their functional three-dimensional forms). A computational grid composed of individual workstations networked together can function as a highly integrated system dedicated to performing a single computational task. Yet it is unproblematic envisioning how the overall system supports decomposition into distinct but interfaced components—the individual workstations. Contra Haugeland, grid interfaces and components situated at those interfaces need not be identified in terms of low-bandwidth bottlenecks. Interface points between nodes in distributed networks of this sort can be as high-bandwidth as is called for to carry out the computational task. Interfaces in distributed computing networks are instead constituted by well-defined points of detachment and reassembly.[11] Therefore, contrary to Haugeland's intuition, there is no intrinsic limitation on interfaces supporting only relatively low-bandwidth couplings between system components.

A second problem for Haugeland's account is the fact that his guiding example is predicated on the assumption that high-bandwidth reciprocal interaction and continuously unfolding feedback loops between motor areas of the brain and the body are the rule and not the exception during normal episodes of sensorimotor behavior. Recent successes with brain-machine interfaces suggest that as an empirical matter sensorimotor control might not require the sort of high-bandwidth coupling between brain and body that Haugeland implies. One recent study, for instance, demonstrated that motor command signals derived from as few as 20–60 neurons within primary motor cortex are capable of supplying the information needed to produce competent and naturalistic reaching behavior with a prosthetic arm (Velliste et al. 2008). This suggests that high-bandwidth connections between motor areas in the brain and the body might not be needed or used after all. Moreover, in virtue of this potential low-bandwidth interface, brain and limb seemingly satisfy Haugeland's own test for being genuine components, since replaceability of the limb with a functional equivalent prosthetic limb is not just in principle but in practice possible.

Third, and more critically, the bandwidth criterion cannot distinguish between high-bandwidth causal coupling among mechanism components from high-bandwidth coupling reflecting causally necessary background conditions of a mechanism's performance (Craver 2007b). The flow of electric current is a causally necessary background condition for a toaster to perform its function, since impeding the flow of current will prevent the toaster from working. These processes are strongly coupled to one another on any available measure of causal strength. Nevertheless, the mechanism supplying the current (the electrical generator) is not part of the toasting mechanism. Similarly, certain membrane proteins use cellular energy (ATP) to actively transport Na+ and K+ ions against their concentration gradients in order to maintain a specific distribution of charged ions across the neuronal membrane and offset the natural movement of ions resulting from

[11] Grush (2003) proposes the *plug points criterion* as a means of delimiting boundaries along similar lines.

diffusion. The activities of these ion pumps, and the intracellular and extracellular ion concentrations they produce, are not generally regarded as part of the action potential mechanism proper. Instead, the sodium–potassium pump is a tightly coupled causal background condition enabling action potential generation. The intuitive judgment returned in these and many other cases is that purely background conditions lie outside the boundaries of these mechanisms. As Craver has highlighted, a key requirement of a mechanism boundary criterion is that it should be capable of sorting high-bandwidth causal coupling between genuine components of a mechanism from comparably high-bandwidth background conditions and activities that merely provide the support structure for a mechanism's normal performance, but are not a part of it. Without additional refinement, the bandwidth criterion is unable to deliver this result.

An interrelated problem identified by Craver is that the bandwidth criterion also cannot distinguish genuine mechanism components, and so mechanism boundaries, from causally coupled activities and processes that play no functional role in the mechanism. These are what Craver (2007b) calls "sterile effects". They are activities or components that, in spite of being engaged when a mechanism is, do not on their own produce changes in the other components of the mechanism or make any real difference to the behavior exhibited by the mechanism as a whole. Changes in neuronal spiking activity associated with information processing in the brain produce predictable changes in local regional blood flow. This strong correlation makes fMRI possible. And yet, the increase in blood flow following activation in the relevant functional areas of the brain is not generally viewed as part of the mechanism for producing action potentials. It is a functionally inert, concomitant effect—a mere correlate. Similarly, heat dissipation and isomerization are both strongly correlated in the retina because photon absorption is their common cause. Most of the time, the energy imparted from an absorbed photon causes conformational changes in the visual pigment molecule rhodopsin, serving to activate the molecule and trigger visual transduction. However, some fraction of photons reaching the retina fail to induce isomerization and instead this energy is dissipated as heat in the surrounding tissue (Rodieck 1998). Consequently, heat dissipation in the retina is a sterile effect of phototransduction. The problem for the bandwidth criterion, or any criterion based on sheer strength of causal interaction, is that they have no resources to disentangle these spurious causal effects from high-bandwidth coupling between mechanism components.

What goes wrong with the bandwidth criterion is its failure to identify, among the multitude of causal relationships in a given system, which are, and which are not, explanatorily relevant to the phenomenon to be explained. The last two problems facing the bandwidth criterion make it particularly clear why mere identification of patterns of intra- and extra-systemic causal interactions alone is insufficient to appropriately demarcate mechanism boundaries. Moreover, they indicate why a further filtering of factors on the basis of their relevance to the overall behavior of a mechanism is essential. The suggestion I now want to pursue is that the boundaries of a given mechanism are to be drawn around all and only the entities, activities, and organizational features explanatorily relevant to the target phenomenon they produce.

The mutual manipulability criterion

All the criteria canvassed above, including the bandwidth criterion, have been relatively remote from considerations biologists and neuroscientists routinely deploy to determine if a given structure or process is mechanistically relevant to producing a behavior or phenomenon of interest. By contrast, Craver (2007a, b) offers a criterion he dubs *mutual manipulability* that is designed to capture the way neuroscientists and biologists experimentally determine mechanism components and boundaries, and how they distinguish genuine components from causal background conditions for a mechanism's normal performance. The mutual manipulability criterion is effectively a boundary demarcation criterion, since the total set of mechanism components determined by the criterion marks the inner-outer boundary of that mechanism. As will be evident below, it is also a *generic* demarcation criterion in the relevant sense that it requires no special assumptions about the nature of cognition. Because intervention-delimited boundaries are resilient to challenges arising from these assumptions, the EC debate can thus be resolved without first settling more controversial debates about the nature of cognition. As indicated above, this is valuable because as long as proposed criteria depend upon prior adoption of some controversial assumption about cognition, the bump in the rug is simply shifted to that domain. If we can solve the problem without such assumptions, we are therefore making considerable progress.

To motivate the mutual manipulability account, consider how neuroscientists typically go about determining components in neural mechanisms underlying cognitive capacities. Back in the 1960s, relatively little was known about the neural basis of tactile discrimination capacities. Mountcastle and colleagues (Talbot et al. 1968) reasoned that neurons in primary somatosensory cortex (S1) might be involved. They probed the specific role of these neurons by applying mechanical vibratory "flutter" stimuli (oscillating at a frequency between 5 and 50 cycles per second) to the skin of awake monkeys while simultaneously recording extracellularly from single units in S1. They observed that a subset of S1 neurons (which they dubbed *quickly adapting* or *QA-neurons*) elicit trains of action potentials precisely time-locked to the oscillations of the vibrating stimulus, suggesting the role of these neurons as components in the mechanism underlying tactile discrimination. In an elegant follow-up experiment undertaken nearly 40 years later, Romo et al. (1998) employed electrical microstimulation to further investigate these putative neural components within S1. Monkeys were first trained to discriminate the frequency of the same "flutter" stimuli applied to skin of the hand, and were rewarded for correctly reporting whether the first or the second stimulus in a test sequence had the higher frequency vibration. Then, one of the two mechanical stimuli was replaced by direct electrical microstimulation of QA-neurons, so that the monkey's new task was to discriminate the frequency of an artificial electrical stimulus delivered to S1 from the frequency of a natural skin vibration. Remarkably, discrimination performance remained unaltered and monkeys continued to perform the frequency discrimination task as if the electrical stimulation were a mechanical stimulus applied directly to the skin. These results, in conjunction with the original correlational findings reported by Mountcastle, were taken to provide strong

evidence that QA-neurons are components in the neural mechanism subserving tactile discrimination. In general, when these two complementary experimental strategies are successfully combined in this manner—engaging subjects in task performance while monitoring changes in putative component(s), and manipulating putative component(s) while detecting changes in overall task behavior—neuroscientists gain confidence that they have discovered the real components in a neural mechanism.

The mutual manipulability account makes explicit the norms embodied in these practices. According to the account, mechanism or system boundaries are determined by relationships of mutual manipulability between the properties and activities of putative components and the overall behavior of the mechanism in which they figure. More specifically, Craver defines the mutual manipulability criterion in terms of two interrelated conditions:

> (M1) When φ is set to the value φ_1 in an (ideal) intervention, then ψ takes on the value f (φ_1) [or some probability distribution of values f (φ_1)].
>
> (M2) When ψ is set to the value ψ_1 in an (ideal) intervention, then φ takes on the value f (ψ_1) [or some probability distribution of values f (ψ_1)] (Craver 2007a, 155–160).

where ψ is a variable[12] standing for some higher-level phenomenon to be explained (e.g., tactile discrimination) and φ is a variable standing for some lower-level component in the mechanism underlying the phenomenon ψ (e.g., spiking activity in S1 neurons).[13] The notion of ideal intervention bears close connections to Woodward's (2003) treatment, and is intended to restrict the kind of intervention specified in M1 and M2 to just those suitably controlled experimental interventions carried out on some variable that can help us to ascertain when changes in its value are in fact causally relevant to changes in another variable of interest. Consider only M1 for the moment. It imposes a restriction such that for a given intervention to change the value of a lower-level component φ, if some change in the value of ψ occurs, it occurs only in virtue of the change in the value of φ and not via some other more indirect route. In particular, Craver has in mind restrictions on interventions which either change the value of ψ directly without involving any changes in φ at all, change the value of ψ indirectly but through some other route besides through φ, or are merely correlated with some other variable that induces

[12] Although manipulability theories of causal explanation are commonly expressed using variables, one should not infer that such accounts are committed to causal relationships holding between "abstracta" rather than objects and properties in the world. See Craver (2007a, 94–95) and Woodward (2003, 14) for further discussion.

[13] The term 'level' here means *mechanistic level*, i.e., the set of component parts and activities responsible for producing a phenomenon or performing some higher-level role (Craver 2007a, b). Lower mechanistic levels bear a compositional relationship to higher mechanistic levels in the sense that lower-level parts are components of the mechanism for a phenomenon at some higher level. A crucial feature of mechanistic levels is that they support recursive decomposition. The activities of components in a mechanism responsible for some phenomenon (constituting one mechanistic level) can themselves be viewed as phenomena to be explained, and still lower-level activities and entities can subsequently be identified in their explanations.

changes in the value of ψ. The same analysis applies to M2. Interventions of this kind place us in a much better position to ascertain something about the componency relationship between the variables of interest. When a lower-level entity or process and higher-level phenomenon sustain relations of mutual manipulability, as expressed by M1 and M2, the entity or process is counted as a component falling within the boundaries of the mechanism for the phenomenon.[14]

As mentioned above, M1 and M2 are designed to embody the evaluative criteria implicit in the experimental strategies that neuroscientists and experimental biologists use to test mechanism componency claims. As should be evident from M1 and M2, Craver maintains that such claims are investigated through *inter-level* experiments involving targeted interventions and detection techniques deployed at different levels of a given mechanism.[15] As he defines them, inter-level experimental interventions test the hypothesized relationship between putative component parts of a mechanism (the entities, activities, and organizational features at some lower level) and the explanandum phenomenon (at some higher level).[16] Inter-level experiments also come in two basic varieties.

"Bottom-up" experiments (captured by M1) involve experimental interventions to induce targeted changes (excitation or inhibition) in the properties and activities of putative lower-level mechanism components while changes are detected in the phenomenon of interest at some higher-level. If the experimental variable subject to intervention is a genuine component in the mechanism, then intervening to change its state (e.g., enhance or suppress its activity) will have some overall effect on the system in which it functions as a component part. Electrical microstimulation, transcranial magnetic stimulation (TMS), and optogenetic control are examples of bottom-up techniques in neuroscience. When changes in task performance are observed in conjunction with the deployment of one of these experimental

[14] One might reasonably wonder about the relevant timescales for assessing relationships of mutual manipulability. Because the mutual manipulability account takes direct guidance from scientific practice, answers about the timescales over which such relationships operate will ultimately be guided and constrained by the relevant science. Since this paper focuses on EC claims in certain sectors of cognitive science and neuroscience, the relevant timescales are primarily organismal ones (e.g., developmental, behavioral, and neural). It is, however, beyond the scope of the paper to defend this proposition in any detail or address how the current account handles relationships that occur more slowly over generational and evolutionary timescales such as cumulative changes in technologies and other forms of cognitive scaffolding in human physical and social environments. I thank an anonymous reviewer for raising this issue.

[15] Bechtel and Richardson (1993), Craver and Darden (2001), and Craver (2002) have also discussed these experimental strategies and the critical role they play in the development of mechanistic explanations.

[16] It should be noted that intra-level interventions (intervention and observation at the same mechanistic level) do sometimes play a limited role in establishing componency claims. They can help to refine our characterizations of the phenomenon to be explained and improve our understanding of the organization of components and their activities within one mechanistic level. Nevertheless, intra-level experiments are primarily useful for testing standard etiological-causal claims (e.g., speeding up reaching movements causes increased endpoint variance; increasing cognitive load increases reaction times; increasing the flow of Na+ into a neuron causes Na+ channels to open, etc.); and inter-level interventions remain the principal means by which componency claims are tested and confirmed. See Craver (2007a, b) for further discussion.

manipulations, a defeasible inference is drawn to the effect that the perturbed entity or structure is a component in the mechanism underlying the phenomenon.

"Top-down" experiments (captured by M2), by contrast, involve interventions to change some aspect of the explanandum phenomenon or overall behavior while changes in the properties and activities at some lower mechanistic level are tracked. In a top-down experiment, one intervenes to engage a subject in some task and monitors for effects on the lower-level properties and activities of the putative components in the mechanism. Neurophysiology and neuroimaging paradigms commonly employ this strategy, engaging animals or human subjects in cognitive tasks (e.g., navigating a maze, planning a reach, or discriminating the speed and direction of visual motion), while changes in neural activity in single cells, small populations, or entire brain regions are monitored.

To be clear, producing the appropriate pattern of changes from either bottom-up or top-down interventions alone provides preliminary but insufficient evidence that the altered element is in fact a component in the mechanism for the target phenomenon. Instead, appropriate evidence from *both* types of intervention must be combined in order to substantiate claims about componency relationships. It is also important to note that the mutual manipulability criterion only states a sufficient condition to establish something as a component part of a mechanism. Because the conjectured criterion does not aim to outline necessary conditions for something to count as a component, there are substantial limitations placed on the conclusions that can be drawn from failures to meet it. Most importantly, one cannot conclude from a failure to satisfy this criterion that a putative component is not in fact a component.

The key advantage of the mutual manipulability criterion over other demarcation criteria including the bandwidth criterion is its ability to distinguish sterile effects, mere correlates, and background conditions, one the one hand, from relevant lower-level components for a phenomenon, on the other. First, consider sterile effects and correlates. The answer returned here is straightforward. One cannot change the phenomenon or the behavior of the mechanism as a whole by intervening to excite or inhibit mere lower-level correlates or sterile effects. However, one can change the behavior of the mechanism as a whole by intervening to manipulate lower-level components. As described above, performance of a cognitive task is well correlated with hemodynamic changes, but this does not mean that the hemodynamic changes are part of the mechanism involved in task performance. Few, if any, cognitive neuroscientists think this. Instead, they view the hemodynamic response as a useful surrogate MRI signal reliably correlated with underlying neural activity (e.g., Logothetis 2008). Appeals to manipulability relations can thus rule out hemodynamic changes as components.[17]

[17] One might object that preventing blood flow to a region would quickly degrade task performance, perhaps along with other long-term consequences, and thus this process should count as a component. However, because regional increase in cerebral blood flow temporally lags behind neural activation by some small amount, it is safe to assume that preventing those changes cannot, strictly speaking, alter either neural activation or the task performance it supports. An ideal intervention to selectively suppress only the time-lagged hemodynamic response following activation would demonstrate this.

How can the mutual manipulability criterion help with the issue of locating high-bandwidth causal background conditions outside the boundaries of cognitive mechanisms? If anything in the local environment plays the role of causal background condition for normal human cognitive, perceptual, and sensorimotor capacities, it is such things as having air to breath with the right ratio of oxygen to carbon dioxide; ready supplies of nutrients from the environment to power and sustain vital metabolic, respiratory, and circulatory processes; and even having background gravitational forces be neither too strong nor too weak for motor behavior to be possible at all. For instance, Fisk et al. (1993) investigated the effects of background gravitational forces on human sensorimotor control by training subjects to make reaching movements to targets in the normal background 1G environment. After training, subjects were asked to perform the same reaches in different force conditions: a high-force condition in which gravitational force was approximately twice the normal level, and a low-force condition in which it was approximately half the normal level. Clearly this is an instance of a bottom-up experimental intervention on the putative component variable (expressed by M1). Determinate effects on movement trajectories and endpoint accuracy induced by altered gravity were observed. However, this alone does not establish that the background 1G environment should be included as a mechanism component. According to mutual manipulability, if it were functioning as a genuine component, then a top-down experimental intervention to engage a subject in a reaching task should produce observable changes in gravitational force levels (as required by M2). The implausibility of this outcome vindicates the mutual manipulability criterion. It provides principled grounds for the intuitive negative judgment that such factors should not be countenanced as genuine mechanism components, but rather should be classified as causal background conditions for normal mechanism performance. It is exactly these differences that are critically important for placing such factors outside of the boundaries of the mechanism, despite the high-bandwidth coupling they might enjoy with genuine mechanism components.

Sometimes the situation is more complex, and requires conjoining the experimental strategies of interference and stimulation one to out the genuine components of a mechanism from background conditions. Intervening to inhibit a background condition should reliably inhibit or alter the phenomena (after all, it is a causally necessary background condition). However, one cannot stimulate or elevate the background condition and thereby change the phenomenon exhibited by the mechanism. Adequate oxygen supply is one particularly obvious background condition vital for normal cognitive function, since brain oxygen consumption accounts for nearly twenty percent of total body oxygen consumption. According to mutual manipulability, both M1 and M2 must be satisfied for either the heart or lungs are to be counted as components of cognition. Intervening to shut down either mechanisms of respiration (lungs) or circulation (heart) will interfere with, among very many other things, the neural systems underlying visual object recognition capacities. Despite this, stimulating the lungs or heart does not correspondingly alter visual object recognition capacities, all other things being equal. Thus, the requirement for bottom-up interventions is incompletely satisfied, and the background condition can be identified as such and ruled out as a component of the cognitive mechanism.

The mutual manipulability criterion thus supplies an objective basis from which to distinguish background conditions from components in a cognitive mechanism, and represents a genuine advance over the bandwidth criterion. As a final step in making a case for mutual manipulability as a superior criterion for demarcating cognitive mechanism boundaries, I now show how the mutual manipulability criterion classifies putative examples of extended cognition, and how it can be used to evaluate the conclusions that advocates of these views intend to draw.

Mutual manipulability applied: bodily extension

First, consider an example of a hypothesis about cognition extending into bodily processes. To date, there remains considerable controversy about how the brain produces accurate, goal-directed movements such as a reach to a viewed object. More specifically, there is an ongoing debate about how detailed motor plans must be in order to produce such movements, and whether any of this internal computation can be "offloaded" into the body. According to one dominant proposal heavily influenced by control theory, the brain computes an internal representation (called an inverse model) that maps the state of the limb at each successive time point along the desired movement trajectory into motor commands (e.g., Shadmehr and Wise 2005; Wolpert and Ghahramani 2000). This set of commands can then be passed to a controller that executes the movement through feedforward and feedback control mechanisms. The key element in this model is that the brain directly specifies a detailed motor plan (i.e., the precise sequence of motor commands needed to activate the muscles) to bring about the intended movement at each time step.

By contrast, an alternative proposal, the equilibrium point control hypothesis (Feldman and Levin 1995; Polit and Bizzi 1978), suggests that the brain only serves to specify the endpoint of the movement and instead relies heavily on peripheral reflex loops and the spring-like, elastic properties of the muscles themselves to determine the actual motor output for the entire movement trajectory. According to this model, each set of muscle activations defines a stable equilibrium position for the hand (or other limb) in space, and desired trajectories are achieved through a sequence of shifting equilibrium points that take the limb from its initial position to the specified endpoint. When the limb gets displaced from its current equilibrium point, passive forces tend to return it to that state. To generate a reach to some location *r*, for example, the brain only needs to compute in advance the joint angles (and corresponding muscle activations) for the final equilibrium point of the limb at location *r*. Passive restoring forces produced by the muscles do the rest to bring the limb to this final equilibrium point, and determine that it will pass through a sequence of intermediate configurations that define additional equilibrium points for the system. For equilibrium point control, one could imagine that the hand is literally attached to a control point (representing the target location) by a spring (representing the limb muscles) which moves along the desired trajectory and drags the arm behind it. For the limb to follow along the desired path, however, limb stiffness (defined as the ratio of change in force to change in muscle length) must be

high.[18] If stiffness is low, akin to a spring having the stiffness of a slinky, the hand path achieved would be radically different.

According to equilibrium point control, movement trajectories can be achieved without reliance on overly complex neural computation because a significant part of this process can be "offloaded" or literally implemented in bodily properties of the limbs and musculature. Instead of viewing the brain as computing detailed motor plans to control movements, it is instead assigned the relatively restricted function of modulating factors such as stiffness. Setting aside whether the equilibrium point control hypotheses is well supported empirically, it stands as a clear example of an extended mechanism claim. According to the proposal, some components of the motor control mechanism are internal components located in the brain. However, other components are external to the boundaries of the brain, in so far as they implemented in material properties and dynamics of the agent's body.

Crucially, the mutual manipulability criterion can help to determine whether bodily properties such as muscle stiffness can justifiably be included as proper component parts of the motor control mechanism. According to the criterion, intervening in a top-down fashion to engage a human subject in a sensorimotor control task should produce detectable effects in the activity or functioning of the underlying component. Additionally, intervening on the putative component should produce a corresponding effect in the target behavior or performance to be explained. When a subject performs the motor behavior in the task context, engagement or recruitment of the posited mechanism components—the limb musculature and peripheral spinal-cord-mediated feedback loops—will occur. Therefore, the top-down manipulation requirement is trivially satisfied. According to the theory, intervening to change the state of the purported non-neural component—stiffness of the limb musculature—will change the relationship between the arm's equilibrium point and the movement trajectory that ensues. Thus, an ideal bottom-up intervention on the lower-level component predicts a higher-level effect on the phenomenon as a whole. The request for a bottom-up manipulation is therefore also satisfied in principle. The mutual manipulability criterion thus clearly outlines a path to empirically vindicate the hypothesis that cognitive boundaries extend into the non-neural body. It does so because it provides a concrete, objective basis for asserting that the body plays the role of a real component in the motor control mechanism, rather than merely serving as a causally coupled background condition.

One might object that the preceding example, while sufficing as an example of extended motor control, somehow still falls short as an example of extended *cognition*. Consider, then, a study conducted by Ballard et al. to test the role of saccadic eye movements while subjects performed a memory-demanding copying task (Ballard et al. 1995; also described in Clark 2008). Subjects were asked to copy a pattern of colored blocks (the "model") appearing on a computer monitor using a cursor to select similar blocks from a resource area and assemble them in a workspace. No other restrictions were imposed on subjects, except to carry out the

[18] Muscles, like springs, vary in stiffness. Applying the same load force to two springs of varying thicknesses will produce different increases in spring length directly proportional to their thickness (i.e., a thick spring will increase its length less than a thin one).

task as quickly and accurately as possible. Under the grip of traditional, information-processing models in psychology, subjects might be expected to complete the task in the following way: look at the model, determine which block to move next, hold both color and location information about the item in working memory, get the matching block from the resource area, and place it in the workspace to match the model pattern. It turns out that subjects do not use this memory-intensive strategy. Instead, they make repeated saccades to and from the model, both before and *after* picking up a block—many more eye movements than would be predicted based on the memory strategy described above. The fact that subjects initially store only the relevant information about color needed to select the appropriate block from the resource area, and then move their eyes back to the model to acquire information about where to place the block in the workspace, indicates they deploy memory resources sparingly and as needed to perform the task. Instead of using intensive computation or memory resources, subjects appear to use embodied skills such as eye movements to acquire the relevant pieces of information needed to accomplish the task.

Theorists wanting to make use of this case in support of EC might invoke mutual manipulability to help clarify the claim that saccadic eye movements (and the oculomotor system) are genuine components underlying performance of this cognitive task, rather than simply serving as a mere background condition. How might this work? The top-down experiment has already been described. By engaging subjects in this cognitively demanding task, subjects initiate a characteristic pattern of saccadic eye movements. They also make limb movements, which naturally seem like background conditions and not working parts of the memory mechanism underlying task performance. However, this is where the other branch of mutual manipulability becomes relevant, and the outcome of a bottom-up experimental intervention must be ascertained. It turns out that this was exactly the control experiment Ballard et al. conducted. In the control experiment, subjects performed the task while maintaining gaze fixation at the center of the screen. Because eye movements were effectively eliminated during task performance, the demand for manipulation of the target phenomenon through a bottom-up inhibitory experiment is satisfied. Crucially, for these subjects, time to task completion—one reasonable metric for performance in the task—was roughly three times longer than when eye movements were allowed.[19] Consequently, mutual manipulability appears to be satisfied in this case, and thus the EC proponent has an objective basis from which to claim that saccadic eye movements function as a component in the mechanism underlying cognitive task performance.

Mutual manipulability applied: environmental extension

Before applying the mutual manipulability criterion to one of the hallmark examples of EC, consider how it can help to address the original gecko example. Recall that

[19] It should be noted that additional control experiments were conducted to rule out alternative explanations for degraded performance such as lower visual resolution across the entire workspace due to enforced central fixation. See Ballard et al. (1995) for further details.

example concerned the possibility of an environmentally-extended mechanism underlying gecko gripping abilities that included the local environmental climbing substrate as a component. Although the ability in question is not a cognitive one, the example is still useful to reconsider in the light of mutual manipulability precisely because it involves a claim of genuine environmental extension.

According to the proposal, the climbing substrate is not merely causally contributing to the measured adhesive force profile of the gecko as a necessary background condition, but is instead contributing as a genuine component part in the mechanism underlying gripping. According to the mutual manipulability criterion, a bottom-up experimental manipulation to alter the properties of the substrate should produce observable changes in the action of the gecko gripping mechanism as a whole. This is precisely the result described in our earlier discussion. Recall that one important piece of experimental evidence for the van der Waals hypothesis over competing hypotheses came from the observation that manipulations to increase the surface energy of the climbing substrate correspondingly increased the overall adhesive force exerted by the whole gecko. Mutual manipulability also implies that if the environmental substrate functions as a component, then top-down experimental interventions to engage the gecko in gripping behavior should have some measurable effect on the substrate itself. However, this is manifestly not the case for the same reason that engaging human subjects in reaching behavior produces no changes in background gravity. By appealing to mutual manipulability, the local environmental substrate is consequently relegated to a causal background condition, despite exhibiting (possibly strong) causal interaction with geckos. Mutual manipulability thus provides an objective, unambiguous basis from which to exclude the substrate as a component in an environmentally-extended adhesion mechanism.

Next consider the application of mutual manipulability to another example of an environmentally-extended mechanism (originally described in Clark 1997) involving the role of the local ocean environment in the swimming capabilities of certain fish species. The maximum speeds in some open water fish such as the bluefin tuna have been observed to be much faster (by a factor of about seven) than those predicted from theoretical estimates based on the maximum amount of power their musculature alone can deliver. As Clark describes, the explanation for this seemingly paradoxical observation is that "the tuna find and exploit naturally occurring currents so as to gain speed, and use tail flaps to create additional vortices and pressure gradients, which are then used for rapid acceleration and turning" (1999, 345). Tuna appear capable of exerting precise and effective control over the water flow around their own bodies. This enables them to extract and exploit energy from local hydrodynamic features such as ocean waves, turbulence, and even their own self-generated wake to improve propulsion. The interesting feature for present purposes is the tuna's finely tuned ability to endogenously produce favorable pressure gradients and then control these gradients to optimize swimming performance. Are these short-lived local environmental features proper components in an extended tuna-local-ocean-environment propulsion mechanism? Or are such features merely causal background conditions? The mutual manipulability criterion can supply a concrete answer.

The bottom-up experimental intervention to alter properties of these vortices and pressure gradients in the immediately surrounding ocean environment will likely

induce a predictable change in overall swimming performance. In fact, by eliminating all such exploitable structures we should predict performance to fall into line with those theoretically calculated estimates of maximum swimming speeds based on muscle power alone. The fact that this result is in close accord with the mutual manipulability criterion, suggesting that we might be dealing with an extended mechanism. However, the top-down experimental manipulation must also be considered. Engaging a tuna in swimming behavior will induce detectable changes in some of the environmental properties, namely locally exploitable vortices and pressure gradients generated by its own bodily motion. Thus, in contradistinction to the conclusions reached above concerning the gecko, here these select, organism-generated, environmental properties seem to satisfy the mutual manipulability criterion. Because of this, talk of an environmentally-extended swimming mechanism with external components is legitimized. To be perfectly clear, this is not equivalent to the claim that *all* hydrodynamic structures in the local ocean environment causally relevant to tuna swimming behavior count as genuine mechanism components. For example, despite being potentially relevant to swimming performance, engaging a tuna in swimming behavior cannot significantly change macroscopic, weather-driven ocean currents, or the wake generated by a ship's propeller. These factors can be relevant to behavioral performance as mere causal background conditions, just as salinity or water temperature being within a biologically safe range is a background condition, but not as genuine components in an extended propulsion mechanism. And because such factors would inevitably fail the test of mutual manipulability, this is as it should be.

Finally, where might a hallmark example of EC properly fit within this spectrum of cases? Does it find support from the mutual manipulability criterion? The example I want to discuss comes from Clark and Chalmers (1998). Since the thought experiment is well-known, I will only sketch the barest details essential for present purposes. Two hypothetical individuals, Inga and Otto, are described as using memory resources to perform the same cognitive task: finding their way to a location in Manhattan. Inga performs this task as normal subjects would. She uses the areas of her brain that support working memory to store, access, and retrieve information to find her way to her destination. Otto, on the other hand, suffers from a memory disorder, and therefore has deficits in some of the normal suite of memory capacities noted above. Consequently, Otto must use a notebook as an active external memory resource to find his way to the destination. As Otto is massively overtrained in using the notebook, he can fluidly encode and access this stored information in functionally similar ways to Inga. Otto's externally-augmented memory abilities are thus taken by Clark and Chalmers to provide support for the EC hypothesis. As Clark (2008) puts it, "the actual local operations that realize certain forms of human cognizing include inextricable tangles of feedback, feedforward, and feed-around loops: loops that promiscuously criss-cross the boundaries of brain, body, and world. The local mechanisms of mind, if this is correct, are not all in the head. Cognition leaks out into body and world" (Clark 2008, xxviii).

The looseness of this description is unfortunate and this, along with many comparable characterizations in the EC literature, has led to unnecessary confusion and debate. We can use the machinery of mutual manipulability to provide more

precise interpretations for putative EC cases. Moreover, defenders of EC should not be hostile to this suggestion. In his most recent (2008) book, Clark prefers to talk about cognitive extension in terms of "supersized mechanisms". If this mechanistic description is legitimate, then we should be able to apply the mutual manipulability criterion to provide an objective means for establishing that components of cognitive mechanisms, and so the boundaries of these mechanisms, extend out into structures in the embedding environment. More specifically, it should be possible to perform top-down and bottom-up experimental interventions in this context to put to empirical test hypotheses concerning supersized mechanisms with environmentally-extended components.

Although framed differently, Clark invokes an explicit prediction that something like a bottom-up intervention will succeed. In defending the claim that Otto's notebook is a component part of the environmentally-extended memory mechanism, Clark states it is "part of the local equipment that plays a causal role in the generation of action. Subtract the notebook encoding and Otto does not go to 53rd Street. Replace it with an encoding that mistakenly indicates 56th street, and Otto ends up there instead" (2008, 79). These are hypothesized (and not real) interventions on the putative lower-level mechanism component. However, in order to rule out the external notebook encoding as merely a causally necessary background condition for memory performance, top-down interventions would be needed as well. Only then should it be considered a genuine component according to mutual manipulability. If we were to perform a top-down activation experiment to engage a subject like Otto in a memory task, would we detect observable changes in the mechanism components? Arguably, in a manner identical to the tuna, engaging Otto in the target behavior would cause him to actively modify and exploit his local environmental resource. For example, we would expect Otto to modify the contents of his external memory store by performing rewrite operations and other transformations when engaged in memory-demanding tasks. It thus seems that the hallmark example satisfies the standards laid down by the mutual manipulability criterion. Of course, this result is provisional. One would ideally like to demonstrate this outcome with suitably controlled experimental interventions of the sort outlined in M1 and M2, rather than through a thought experiment. Nonetheless, this represents an important push in the right direction for the prospects of EC. Mutual manipulability criterion thus lays down a clear gauntlet for advocates of EC. If talk of environmentally-extended mechanisms is to have real purchase, then this general criterion for identifying mechanism components should apply equally well in these novel contexts.

The trouble with many claims about cognitive extension is that insufficient attention has been paid to the important difference between causally coupled background conditions and component parts of a given extended cognitive mechanism—a difference I have been at pains to make clearer in this paper. Clark's (1997) description of the multiplicity of neural, bodily, and environmental factors that jointly conspire to produce human infant motor learning is representative: "[T]he developmental pattern is not the expression of an inner blueprint. Rather, it reflects the complex interplay of multiple forces, some bodily (leg mass), some mechanical (leg stretching and spring-like actions), some fully external (the presence of treadmills, water, etc.), and some more cognitive and internal (the transition to volitional—i.e.,

deliberate—motion). To focus on any one of these parameters in isolation is to miss the true explanation of developmental change, which consists in understanding the interplay of forces in a way that eliminates the need to posit any single controlling factors" (1997, 42).

The trouble with this formulation should be entirely obvious now. The point at issue throughout this paper has been the fact that among the manifold causally relevant factors for normal cognitive, perceptual, and motor performance (and their normal developmental patterns), only some are correctly described as components in the mechanism of that performance. Others are more appropriately described as causal background conditions for that performance. Clark here describes the various factors germane to explaining how human infants learn to walk in a way that is entirely insensitive to this distinction. As I have argued, there are objective considerations embodied in the mutual manipulability criterion that can be brought to bear in order to segregate these factors appropriately. Clark and numerous others have undoubtedly been helpful in expanding the scope of candidate factors potentially relevant to explaining human cognition. However, what these previous efforts have failed to do is state in a clear manner exactly why a subset of these factors should be recognized as genuine mechanism components rather than mere causal background conditions. I have suggested in this paper that, to date, the EC camp has been unable to supply such a defense. Until they do, this will stand as the single greatest challenge to winning broader acceptance within the community of cognitive scientists and scientifically oriented philosophers who care about such matters.

Conclusions

The mutual manipulability criterion performs better than other extant criteria at drawing the appropriate boundaries around those components responsible for producing the target phenomenon. What counts as a genuine component of a mechanism (cognitive or otherwise)—and so what is located within the boundaries of a mechanism and what is outside—is determined by the presence of relationships of mutual manipulability between the properties and activities of putative components and the overall behavior of the mechanism in which they figure. Among its main advantages as a criterion for demarcating cognitive boundaries, the mutual manipulability criterion stands alone in its ability to distinguish the components underlying cognition from causally coupled background conditions. Moreover, an intervention-based criterion establishes a direct link to the primary methods by which cognitive scientists and neuroscientists empirically test and determine mechanism boundaries. Thus, by using the mutual manipulability criterion, the philosophical debate over EC is brought into closer contact with the relevant sectors of cognitive science and neuroscience. Finally, some widely discussed cases in the EC debate successfully meet the standards laid out by mutual manipulability, demonstrating its direct applicability and usefulness in this context. This result indicates that conjectures about bodily- and environmentally-extended mechanisms can be tested empirically and justified on the basis of inter-level experiments involving mutual manipulability relationships. Proponents of EC would therefore be well advised to

determine how their own proposals square with respect to the mutual manipulability criterion. Otherwise, the burden of proof falls upon them to provide arguments as to why this generic criterion for demarcating mechanism boundaries, which finds wide application across the biological and cognitive sciences, is inapplicable or irrelevant in the context of extended cognitive mechanisms.

Acknowledgments Thanks to Jake Beck, Carl Craver, Philip Gerrans, Peter Langland-Hassan, Gerard O'Brien, and Gualtiero Piccinini for helpful comments on previous drafts of this paper. Thanks also to an anonymous reviewer and the editor at the journal for constructive feedback.

References

Adams F, Aizawa K (2001) The bounds of cognition. Philos Psychol 14:43–64
Adams F, Aizawa K (2008) The bounds of cognition. Blackwell, Oxford
Adams F, Aizawa K (2010) Defending the bounds of cognition. In: Menary R (ed) The extended mind. Ashgate, Aldershot
Autumn K (2006) How gecko toes stick. Am Sci 94:124–132
Autumn K, Liang YA, Shie ST, Zesch W, Chan WP, Kenny TW, Fearing R, Full RJ (2000) Adhesive force of a single gecko foot-hair. Nature 405:681–684
Ballard D, Hayhoe M, Pelz J (1995) Memory representations in natural tasks. Cogn Neurosci 7:66–80
Bechtel W (2008) Mental mechanisms: philosophical perspectives on cognitive neuroscience. Routledge, London
Bechtel W, Richardson RC (1993) Discovering complexity: decomposition and localization as strategies in scientific research. Princeton University Press, Princeton
Clark A (1997) Being there: putting brain, body, and world together again. MIT Press, Cambridge
Clark A (1999) An embodied cognitive science? Trends Cogn Sci 3(9):345–351
Clark A (2005) Intrinsic content, active memory, and the extended mind. Analysis 65(10):1–11
Clark A (2007) Curing cognitive hiccups: a defense of the extended mind. J Philos 104(4):163–192
Clark A (2008) Supersizing the mind: embodiment, action, and cognitive extension. Oxford University Press, New York
Clark A, Chalmers DJ (1998) The extended mind. Analysis 58:10–23
Craver CF (2002) Interlevel experiments and multilevel mechanisms in the neuroscience of memory. Philos Sci Suppl 69:S83–S97
Craver CF (2007a) Explaining the brain. Oxford University Press, New York
Craver CF (2007b) Constitutive explanatory relevance. J Philos Res 32:3–20
Craver CF, Darden L (2001) Discovering mechanisms in neurobiology: the case of spatial memory. In: Machamer PK, Grush R, McLaughlin P (eds) Theory and method in neuroscience. University of Pittsburgh Press, Pittsburgh, pp 112–137
Dennett DC (1987) The intentional stance. MIT Press, Cambridge
Feldman AG, Levin MF (1995) Positional frames of reference in motor control: origin and use. Behav Brain Sci 18(4):723–806
Fisk J, Lackner JR, Dizio P (1993) Gravitoinertial force level influences arm movement control. J Neurophysiol 69(2):504–511
Fodor JA (2009) Where is my mind? Review of supersizing the mind: embodiment, action, and cognitive extension. Lond Rev Books 31(3):13–15
Grush R (2003) In defense of some 'Cartesian' assumptions concerning the brain and its operation. Biol Philos 18:53–93
Haugeland J (1995/1998) Mind embodied and embedded. Acta Philosophica Fennica 58:233–267 (Reprinted in Haugeland, J. Having Thought. Cambridge, MA: Harvard University Press)
Logothetis NK (2008) What we can and what we cannot do with fMRI. Nature 453:869–878
Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. Philos Sci 67:1–25
Nelson WL (1983) Physical principles for economies of skilled movements. Biol Cybern 46:135–147
Polger T (2004) Natural minds. MIT Press, Cambridge
Polit A, Bizzi E (1978) Processes controlling arm movements in monkeys. Science 201(4362):1235–1237

Putnam H (1975) The nature of mental states. In: Mind, language, and reality. Cambridge University Press, Cambridge
Rodieck RW (1998) The first steps in seeing. Sinauer, Sunderland
Romo R, Hernández A, Zainos A, Salinas E (1998) Somatosensory discrimination based on cortical microstimulation. Nature 392(6674):387–390
Rowlands M (1999) The body in mind. Cambridge University Press, Cambridge
Rowlands M (2010) The new science of the mind: from extended mind to embodied phenomenology. MIT Press, Cambridge
Rumelhart DE, Smolensky P, McClelland JL, Hinton G (1986) Schemata and sequential thought processes in PDP models. In: McClelland JL, Rumelhart D (eds) Parallel distributed processing: explorations in the microstructure of cognition, vol. 2: psychological and biological models. MIT Press, Cambridge, pp 7–57
Rupert R (2004) Challenges to the hypothesis of extended cognition. J Philos 101(8):389–428
Searle J (1980) Minds, brains, and programs. Behav Brain Sci 3:417–424
Shadmehr R, Wise SP (2005) The computational neurobiology of reaching and pointing: a foundation for motor learning. MIT Press, Cambridge
Shapiro LA (2009) Review of the bounds of cognition. Phenomenol Cogn Sci 8:267–273
Simon HA (1969) The sciences of the artificial. MIT Press, Cambridge
Talbot WH, Darian-Smith I, Kornhuber HH, Mountcastle VB (1968) The sense of flutter-vibration: comparison of the human capacity with response patterns of mechanoreceptive afferents from the monkey hand. J Neurophysiol 31(2):301–334
Velliste M, Perel S, Spalding MC, Whitford AS, Schwartz AB (2008) Cortical control of a prosthetic arm for self-feeding. Nature 453:1098–1101
Wheeler M (2010) Extended functionalism. In: Menary R (ed) The extended mind. MIT Press, Cambridge
Wilson RA (2004) Boundaries of the Mind: the individual in the fragile sciences. Cambridge University Press, Cambridge
Wolpert DM, Ghahramani Z (2000) Computational principles of movement neuroscience. Nat Neurosci 3:1212–1217
Woodward J (2003) Making things happen: a theory of causal explanation. Oxford University Press, New York