4 Defending the Bounds of Cognition

Fred Adams and Ken Aizawa

Introduction

Question: Why did the pencil think that 2 + 2 = 4?
Clark's answer: Because it was coupled to the mathematician.

That about sums up what is wrong with Clark's extended mind hypothesis. Clark apparently thinks that the nature of the processes internal to a pencil, Rolodex, computer, cell phone, piece of string, or whatever, has nothing to do with whether that thing carries out cognitive processing.[1] Rather, what matters is how the thing interacts with a cognitive agent; the thing has to be coupled to a cognitive agent in a particular kind of way. Clark (this volume) gives three conditions that constitute a rough or partial specification of the kind of coupling required:

1. The resource has to be reliably available and typically invoked.
2. Any information retrieved from/with the resource must be more or less automatically endorsed. It should not usually be subject to critical scrutiny (unlike the opinions of other people, for example). It should be deemed about as trustworthy as something retrieved clearly from biological memory.
3. Information contained in the resource should be easily accessible as and when required (Clark, this volume, p. 46).

Granted condition 3 doesn't fit the use of a pencil very well, since the mathematician is not really extracting information from the pencil, but blame Clark for that. After all, he likes the idea that the use of pencil and paper in computing sums constitutes part of an agent's cognitive processing; hence it's up to him to make his story work there.[2]

When Clark makes an object cognitive when it is connected to a cognitive agent, he is committing an instance of a *coupling-constitution fallacy*.

This is the most common mistake that extended mind theorists make.[3] The fallacious pattern is to draw attention to cases, real or imagined, in which some object or process is coupled in some fashion to some cognitive agent. From this, one slides to the conclusion that the object or process constitutes part of the agent's cognitive apparatus or cognitive processing. If you are coupled to your pocket notebook in the sense of always having it readily available, use it a lot, trust it implicitly, and so forth, then Clark infers that the pocket notebook constitutes a part of your memory store. If you are coupled to a rock in the sense of always having it readily available, use it a lot, trust it implicitly, and so forth, Clark infers that the rock constitutes a part of your memory store. Yet coupling relations are distinct from constitutive relations, and the fact that object or process *X* is coupled to object or process *Y* does not entail that *X* is part of *Y*. The neurons leading into a neuromuscular junction are coupled to the muscles they innervate, but the neurons are not a part of the muscles they innervate. The release of neurotransmitters at the neuromuscular junction is coupled to the process of muscular contraction, but the process of releasing neurotransmitters at the neuromuscular junction is not part of the process of muscular contraction. (That's a quick and dirty run through the coupling-constitution fallacy. For a less quick and dirty treatment, see Adams and Aizawa 2008.)

So, if the fact that an object or process *X* is coupled to a cognitive agent does not entail that *X* is a part of the cognitive agent's cognitive apparatus, what does? The nature of *X*, of course. One needs a theory of what makes a process a cognitive process rather than a noncognitive process. One needs a theory of the "mark of the cognitive." It won't do simply to say that a cognitive process is one that is coupled to a cognitive agent, since this only pushes back the question. One still needs a theory of what makes something a cognitive agent. This is another weakness of extended mind theories. Yet, in all fairness to Clark and other extended mind theorists, it must be admitted that one of the shortcomings of contemporary cognitive psychology is that there is no well-established theory of just exactly what constitutes the cognitive. Be this as it may, Adams and Aizawa (2001) set out a rather familiar proposal, namely, that cognition is constituted by certain sorts of causal processes that involve nonderived content. We motivated this proposal in two ways, by appeal to examples in other sciences, such as chemistry and physics, and by appeal to what appear to be psychological laws. We mentioned in particular psychophysical laws, such as Weber's law, and psychological laws governing memory formation and recall. We might well have extended our examples by appeal to further

examples to be found in cognitive psychology textbooks. What we, therefore, proposed is that the weight of empirical evidence supports the view that, as a matter of contingent empirical fact, there are processes that (a) are recognizably cognitive, (b) take place in the brain, (c) do not take place outside of the brain, and (d) do not cross from the brain into the external world.

We think that Clark has not yet come to grips with what we are getting at with the view that cognition is a species of causal processing involving nonderived content. Our paper did not provoke him to address what seems to us to be the two most widespread problems with extracranial and transcranial theories of tool use. That is to say, Clark provides no response to the coupling-constitution fallacy, and he provides little more than a hint at what *he* thinks distinguishes the cognitive from the noncognitive. Further, we are disappointed that we were unable to convey our objections clearly enough to forestall Clark's criticisms.

1 The Intrinsic Content Condition[4]

In Adams and Aizawa 2001, we proposed that "A first essential condition on the cognitive is that cognitive states must involve intrinsic, non-derived content" (p. 48). This hypothesis has some calculated openness in it.[5] Suppose that during the course of a cognitive process an agent entertains the thought that John loves Mary. This cognitive agent might thus pass through a cognitive state containing the representation JOHN LOVES MARY. Then, our proposed condition would be satisfied. But, suppose that instead the cognitive agent passed through a cognitive state that has JOHN LOVES MARY followed by a period or maybe some parentheses thrown in. Still, our proposed condition on the cognitive would be satisfied. The hypothesis has this latitude, since we think that although we have good reasons to believe in the existence of intrinsic content, we have no good reasons to think that cognitive states must consist entirely of intrinsic representations or that cognitive states must be, in their entirety, content bearing.[6] This is why we said that "it is unclear to what extent each cognitive state of each cognitive process must involve non-derived content" (Adams and Aizawa 2001, p. 50).

Despite our attempts to present the foregoing position clearly, Clark criticizes us both for being too demanding and too lenient on the role we think nonderived content plays in cognition. Early in his section on intrinsic content he writes, "The question is, must everything that is to count as part of an individual's mental processing be composed solely and exclusively of states of affairs of this latter intrinsically content-bearing

kind? I see no reason to think that they must" (Clark, this volume, p. 48). Here Clark tars us with the overly strong view which we explicitly rejected, then proceeds to critique the overly strong view. (We shall return to this critique, which we find unconvincing.) Later, when Clark comes to our claim about the extent to which each cognitive state of each cognitive process must involve nonderived content, he treats the qualification as rendering the condition vacuous. But this is not a very serious attempt to understand what we are after. Clearly, we mean that if you have a process that involves no intrinsic content, then the condition rules that the process is noncognitive. In fact, that is exactly what the condition is used to show in our 2001 essay. The images on the CRT screen of the Tetris video game are not representations of blocks to be rotated; they are the blocks to be rotated.[7]

Although Clark attributes to us a view we rejected, we find that his case against this misinterpretation is unconvincing. We want to review this here simply to clarify, where we can, features of the distinction between derived and nonderived content. So, what is Clark's case against thinking that not all of an individual's cognitive states must be exhaustively constituted by nonderived representations? It is the following:

> suppose we are busy (as part of some problem-solving routine) imaging a set of Venn diagrams/Euler circles in our mind's eye. Surely the set-theoretic meaning of the overlaps between, say, two intersecting Euler circles is a matter of convention. Yet this image can clearly feature as part of a genuinely cognitive process. (Clark, this volume, p. 48)

Evidently the problem here is supposed to be that there are some mental states that have contents in virtue of a social convention. So, Clark implies that there are bona fide cognitive processes that involve derived content. Clark explores a line of response he thinks we might try. That line, however, strikes us as very weak. We'll bother with none of it. Our view is that Clark's analysis of the Euler circles case is superficial and confused.

To begin, let us draw a rough-and-ready distinction between mental representations of natural objects and mental representations of objects with derived content.[8] The idea is that there are mental representations of things like trees, rocks, birds, and grass, on the one hand, and mental representations of words, stop signs, warning lights, and gas gauges, on the other. Perhaps a better terminology can be chosen, but the names are really inessential. By our lights, words, stop signs, warning lights, and gas gauges mean what they do through some sort of social convention. By our lights, mental representations of natural objects, such as trees, rocks, birds, and

grass, mean what they do in virtue of satisfying some naturalistic conditions on meaning. Many of the essays in Stich and Warfield 1994 present some of the options that philosophers have proposed in these latter cases. Clark's example of the Euler circles draws attention to a muddier case, the case of mental representations of items with derived content. How do these get their meanings?

As noted above, Clark suggests that mental representations of items with derived content get their content by social convention. Now, it is common ground that social convention is in some sense involved in the meaning of the overlap of Euler circles. But that is a logically separate matter from what makes an imagistic mental representation of intersecting Euler circles mean what they do. Intersecting Euler circles on paper getting their meaning is one thing; intersecting Euler circles in mental images getting their meaning is another. Clark apparently overlooks this difference, and hence does not bother to provide a reason to think that Euler circles in mental images get their meaning via social convention. For all Clark says, mental items that have Euler circles as their content could mean what they do by some naturalistic theory of content, just as we suppose that mental representations of natural objects do. So, for all Clark says, a mental image of an intersection of two Euler circles means what it does in virtue of satisfying the conditions of Fodor's (1994) asymmetric causal dependency theory of content. Moreover, what we have just said about Euler circles applies just as well to mental representations of words, stop signs, white flags, and warning lights. It can be a matter of convention that "dog" means dog, that a stop sign means that you should stop, that a person raising a white flag means to surrender, and that a flashing red light means that something is overheating. But that does nothing to show that it is not the satisfaction of some set of naturalistic conditions on nonderived content that gets something in the head to have the meanings of "dog," a stop sign, a white flag, and a warning light.

But suppose Clark acknowledges that there is a conceptual difference between how mental objects get their contents and how artifacts outside the mind get theirs. He might give the following argument for his view. He might still think that there cannot be mental images in which intersecting Euler circles mean set-theoretic overlap unless there were a social convention according to which intersecting Euler circles meant set-theoretic overlap. He might say that this is a kind of derivation of meaning. The meaning of the mental image derives in part from the prior existence of the meaning of physical pictures. The meaning of the mental image might be said to depend on the existence of a prior meaning.

At first blush this argument may seem compelling, but in reality the argument merely trades on an ambiguity in the notions of derivation and dependency.[9] Insofar as there must be a social convention regarding the intersections of Euler circles in order to have a mental representation regarding the intersections of Euler circles, this is not a fact about the constitution of the content of a mental image of the intersections of Euler circles. It is, if anything, a kind of historical fact.[10] One would not have a mental image involving the intersection of Euler circles meaning set-theoretic overlap without having had at some prior time the social convention involving the intersection of Euler circles meaning set-theoretic overlap. It is like this: The dependence of meaning of the mental image of intersecting Euler circles on the social contrivance regarding the intersection of Euler circles is just like the dependence of the meaning of a mental representation of a car on the contrivance of a car. Had the car not been invented, there would not have been mental images of cars. Had the usage of Euler circles not been invented, there would not have been mental images of Euler circles for set-theoretic purposes. This sort of historical truth, if it is a truth, does not show what Clark might want it to show, namely, that the content of certain mental items derives (in the relevant sense) from a social convention.

Suppose, now, that Clark concedes that there is a conceptual difference between how mental objects get their meaning and how physical objects outside the mind get their meaning and admits that he has no argument for the former having derived content, but then demands some reason to think that mental objects do not have derived content. Maybe he has no argument in support of his view, but what reason is there against his view? In the arrangement of social conventions, we have some access to the items bearing the content we want. A community might get together and decide that a yellow flag, rather than a white flag, means surrender, that "bad" or "cool" makes a positive commentary on a thing, or that "WC" is a symbol for the facilities. To do these things, there has to be some way to specify or access the would-be syntactic item that is to figure in the semantic convention. Yet, with the brain, we have no such access to the syntactic items we would like to have bear a particular content. We cannot make, say, the firing of a particular set of neurons mean what it does simply by an agreement that it does. We cannot do this because we have no way to identify particular tokens of brain states qua syntactic items in order to affix contents to them. Given the state of current science, we only identify a person's brain states via inferences to the content of those states. We think that Jones wants to go to that restaurant in Philly because she

said she wants to go to that restaurant and is looking up the address in the phone book. Even when we know that Jones wants to go to that restaurant in Philly, we don't know what specific syntactic item in the brain bears that content. This is not how conventional meanings work.

So, as far as we can tell, Clark gives no reason to doubt what we think is false, namely, that all cognitive states must be exhaustively constituted by content-bearing items. Much less does he give any reason to doubt what we think is true, namely, that cognitive states must involve nonderived content. Further, there are reasons to believe that cognitive content is not normally derived via any sort of social convention. Perhaps there are futuristic science-fiction scenarios in which humans have sufficient access to brain states that this situation could change, but then maybe it will be the case that cognitive content can at times be socially controlled. Maybe. After all, can a mental image of Abraham Lincoln really mean George Washington?

2 The Causal Processing Condition

Our appeal to scientific categorization via causal principles is meant to do two sorts of things for us. First, it is supposed to draw attention to what appears to be one of the principal differences between processes that occur in the brain and processes that occur outside of the brain. Second, it is supposed to draw attention to the unruly collection of processes that might fall under the rubric of a would-be "brain-tool science." Although both of these contentions undermine transcranial theories of cognition, Clark directs most of his attention to the second use of the causal processing condition. He thinks that this argument is doubly flawed. We shall address each of these alleged flaws in turn.

The First Flaw

Clark begins his critique with the following:

> The first thing to say in response to all this is that it is unwise to judge, from the armchair, the chances of finding "interesting scientific regularities" in any domain, be it ever so superficially diverse. Consider, for example, the recent successes of complexity theory in unearthing unifying principles that apply across massive differences of scale, physical type, and temporality. There are power laws, it now seems, that compactly explain aspects of the emergent behavior of systems ranging from the size distribution of cities to word-occurrence frequencies to the frequency of avalanches in sandpiles.

> In a similar vein, it is quite possible that despite the bottom-level physical diversity of the processes that write to, and read from, Otto's notebook, and those that write to, and read from, Otto's biological memory, there is a level of description of these systems that treats them in a single unified framework (for example, how about a framework of information storage, transformation, and retrieval?) (Clark, this volume, p. 50)

We find this passage indicative of a number of respects in which we have failed to make our argument sufficiently clear.

Let's begin by clarifying what we take to be the epistemic status of our view. Clark claims that "it is unwise to judge, from the armchair, the chances of finding 'interesting scientific regularities' in any domain, be it ever so superficially diverse." This may be just a generic rejection of anything like "armchair philosophy." We don't endorse armchair philosophy and we don't see that we are guilty of it. We think that the available empirical evidence provides good reason to think that the chances of finding interesting cognitive regularities covering brains and tools is low. Bear in mind that we side with what is by all accounts scientific orthodoxy. Note as well that Clark does not respond to us by marching out an interesting scientific or cognitive regularity we didn't see from our "armchairs."[11] Alternatively, Clark may be giving an argument for the conclusion that it is unwise to judge the chances of finding interesting scientific regularities that might constitute a "brain-tool science." Clark's argument may be that, just as we have found surprising new regularities through complexity theory, so we might find interesting new regularities in "brain-tool science"; perhaps they will be information-processing regularities. This argument, however, is hardly compelling. Are we to think that a judgment is unwise simply because it could be wrong? More compelling would be to argue that a particular judgment is unwise because it flies in the face of weighty empirical evidence. More compelling would be to show us an interesting cognitive brain-tool regularity that we have overlooked. Yet Clark provides no such case.

Think of the foregoing this way. We maintain that the weight of empirical evidence supports the view that there are processes that (a) are plausibly construed to be cognitive, (b) occur within the brain, (c) do not occur outside of the brain, and (d) do not cross the bounds of the brain. One can challenge the evidence and the argumentation, but it is a bit much to suggest, as does Clark, that there is no evidence whatsoever. We are, after all, siding with scientific orthodoxy. Since it is orthodoxy, there is at least some prima facie reason to think it is not scientifically groundless. Further, the fact that it sides with scientific orthodoxy suggests that the posi-

tion is defeasible. So it hardly helps Clark to point out that we could be wrong.

The observation that *it is possible that* there are higher-level information-processing regularities that cross the boundary of the brain does nothing to challenge our position, which is concerned with what the evidence shows. However, let's see what happens if we grant Clark a much stronger premise. Suppose we detach the modal operator. Suppose that there really are information-processing regularities that cross the boundary of the brain.[12] Perhaps processing information is what Clark thinks constitutes the mark of the cognitive, a condition other than being connected to a cognitive agent.[13] Does this much stronger, nonmodal premise suffice to establish that the mind extends beyond the bounds of skin and skull? No. The problem is that the empirical evidence we have indicates that the brain processes information according to different principles than do common brain-tool combinations. Think of consumer electronics devices. We find that DVD players, CD players, MP3 players, tape recorders, caller ID systems, personal computers, televisions, AM/FM radios, cell phones, watches, walkie talkies, inkjet printers, digital cameras, and so forth, are all information processors. The preponderance of scientific evidence, however, indicates that they process information differently than does the brain. That is why, for example, the brain is capable of linguistic processing, whereas these other devices are not. That is why, for example, the brain is capable of facial recognition over a range of environmental conditions, whereas these other devices are not. This is why the brain is crucial for humans' ability to drive cars, whereas these other devices are not. The differences in information-processing capacities between the brain and a DVD or CD player is part of the story of why you can't play a DVD or CD with just a human brain. These differences are part of the reason you need a radio to listen to AM or FM broadcasts. It is these differences that support the defeasible view that there is a kind of intracranial processing, plausibly construed as cognitive, that differs from any extracranial or transcranial processing. This is the first kind of work we take our appeal to causal processing to do.

We appeal to the nature of causal processing to do more work when we observe that consumer electronics devices and other tools differ among themselves in how they process information. DVD players process information differently than do digital cameras. Digital cameras and DVD players process information differently than do FM radios. This, after all, is what differentiates these tools from each other. What information-processing principles do string, a rock, and DVD players have in common?

When we press this point, we suppose that tools constitute an open-ended set of objects. Tools do not constitute a natural kind; tools are, after all, artifacts. It is for this reason that, a would-be brain-tool science would have to cover more than just a multiplicity of causal processes. It would have to cover a genuine motley. A brain-tool science would not have to cover a mere disjunction of things; it would have to cover an open disjunction. In our 2001 paper, we noted the existence of areas of scientific investigation where there was an apparent fragmentation of a domain.[14] The reason, we argued, that brain-tool science will not go the way of these other investigations is that a would-be brain-tool science would have to cover too broad a collection of processes. It would have to cover a motley of processes, not just a multiplicity of processes.

Clark has hinted that information processing constitutes the mark of the cognitive, but we have argued that this is implausible. What, then, of the possibility that Clark thinks that some other higher-level processes constitute the mark of the cognitive? Perhaps the higher-level processes that extend the mind are of some other nature. Okay; but what are these principles and what is the evidence for their existence? Clark gives us no clue. Note as well that it is not enough for Clark to show that "there is a level of description of these systems that treats [intracranial and extracranial processes] in a single unified framework." Physics provides a reasonable approximation to such a thing. Biology and chemistry might also provide levels of description at which there are processes that are continuous across the boundary of the brain. What Clark needs is a *cognitive* level of description of these systems that treats them in a single unified way. That is, he needs a plausible theory of what constitutes the cognitive. That is where our theory of nonderived content and causal processes supports intracranialism.

The Second Flaw

What, now, of the second way in which Clark thinks our appeal to causal processing is doubly flawed? Clark observes that cognition might fragment into a motley of causally distinct processes without even a family resemblance. Perhaps the folk notion of visual processing will break down into two subtypes: visual processing that eventuates in perceptual experiences and visual processing that guides action independently of perceptual experiences. Extrapolating from what Clark writes, we might add that memory might break down into distinct kinds: short-term memory, long-term memory, visual memory, and so on. A folk notion of auditory processing

could fragment into auditory processing and linguistic processing. Olfaction could have a generic smell component alongside a system for processing pheromones. If cognition is a motley, then Adams and Aizawa's standard will judge intracranial cognitive science just as much a bust as a would-be brain-tool science.

To address this objection, we can apply much of what we said above. To begin with, we do not suppose that the decomposition of the cognitive into a motley is in any sense impossible. We made this epistemic point above. We think that the weight of argumentation supports our view. So, insofar as Clark cares to address our position, he evidently needs at least the nonmodal conclusion that cognition fragments into a motley collection of principles. This, however, we are not prepared to concede. In our earlier discussion we drew a distinction between a multiplicity of principles being at work in some domain and a genuinely motley, open-ended collection of principles being at work. We think that the available scientific evidence makes it plausible that there are distinct sorts of cognitive processing occurring in the brain: processing corresponding to many distinct forms of visual processing, memory processing, and so forth. Yet, we see no reason to extrapolate to the conclusion that there is an open-ended collection. The brain is at least in the running to be a natural kind, whereas brain-tool combinations are hybrids of natural kinds and artifacts. Outside the realm of science fiction, the brain is constrained to develop only a limited set of distinct structures with a bounded range of plasticity. An organism's genome and environmental interactions limit what can be done with neurons and glial cells. Clark appeals to the wide diversity of organisms that might be capable of cognitive processing, but this does not show that there is an open-ended range of things that can constitute cognitive processing. By contrast, tools can be made of anything and can work according to any number of distinct principles. They are clearly artifacts and not natural kinds. That is good grounds for saying that intracranial processing is a collection of disparate mechanisms, whereas brain-tool combinations form an open-ended collection.

Finally, suppose that Clark is right about cognition breaking down into a genuinely open-ended collection of principles. Even that would not necessarily vindicate extracranialist or transcranialist theories of cognition. As long as the multiplicity or motley collection of plausibly cognitive intracranial causal processes is distinct from the set of extracranial and transcranial processes, there will be a basis on which to say that cognition is intracranial. Even if we were to concede the idea that there could be a

science of the motley, a science of the motley would not vindicate extracranialism. So, as far as we can tell, Clark has said nothing that challenges our original analysis of the role of causal processing and nonderived content in the demarcation of the cognitive.

Conclusion

In our essay "The Bounds of Cognition" we thought that the principal weakness in extracranialist theories of tool use was inadequate attention to the mark of the cognitive. Since then, however, we have been impressed with the extent to which this inattention appears to have been involved in so many process externalists' succumbing to one or another version of the coupling-constitution fallacy. It would certainly do much to advance the transcranial theories of cognition were Clark not only to address our theory of the mark of the cognitive, but to address the pervasive coupling-constitution fallacy and set out a plausible theory of what distinguishes the cognitive from the noncognitive.

Notes

1. Clark does shy away from this from time to time, but more on this below.

2. Cf. Clark and Chalmers, this volume, p. 28; Clark 2001, pp. 133–134.

3. Van Gelder and Port (1995), Clark and Chalmers (1998, this volume), Clark (2001), Gibbs (2001), and Haugeland (1998) all make this mistake in one way or another.

4. In a conference presentation in which he responds, in part, to Adams and Aizawa 2001, Clark alludes to Dennett 1990 as providing an argument against nonderived content. Clark does not refer to this argument in this volume, so we have produced an independent critique of Dennett's paper in Adams and Aizawa 2005.

5. See Adams and Aizawa 2001, pp. 50–51.

6. If you think that a cognitive state is a total computational state of a computer, such as a Turing machine, then you will have another reason to doubt the view that a cognitive state must be representational in its entirety. In such views of cognition, at least some of the program states are not representational. That is, for at least some Turing machines, the read-write head of a Turing machine in state S_0, or whatever, is not representational.

7. See Adams and Aizawa 2001, p. 54.

8. We might run what follows using a different terminology. We might talk about states in which the contents are natural objects and states in which the contents

are objects with derived content. We choose to write about mental representations simply for convenience.

9. See Dennett 1990, and our discussion of it in Adams and Aizawa 2005, for another instance of this kind of problem.

10. There is room here to challenge the historical claim that had the use of Euler circles not been invented, there would not have been the use of the mental images of Euler circles. For present purposes, however, we will not pursue this.

11. In truth, when Clark starts "pumping intuitions" (p. 44), talking about Martians (p. 44), and drawing attention to what could happen in science (p. 50), it begins to sound as if he is the one doing armchair philosophy.

12. This is what Rowlands (1999) clearly thinks constitutes a basis for a version of the extended mind hypothesis.

13. This harks back to our opening paragraph.

14. Adams and Aizawa 2001, pp. 60–61.

References

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

Adams, F., and Aizawa, K. (2005). Defending non-derived content. *Philosophical Psychology*, *18*, 661–669.

Adams, F., and Aizawa, K. (2008). *The Bounds of Cognition*. Oxford: Blackwell.

Clark, A. (2001). Reasons, robots, and the extended mind. *Mind and Language*, *16*, 121–145.

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

Dennett, D. (1990). The myth of original intentionality. In K. A. Mohyeldin Said, W. H. Newton-Smith, R. Viale, and K. V. Wilkes (eds.), *Modeling the Mind*. Oxford: Oxford University Press.

Fodor, J. (1994). A theory of content, II: The theory. Reprinted in Stich and Warfield 1994 (pp. 180–222).

Gibbs, R. W. (2001). Intentions as emergent products of social interactions. In Bertram F. Malle, Louis J. Moses, and D. A. Baldwin (eds.), *Intentions and Intentionality* (pp. 105–122). Cambridge, MA: MIT Press.

Haugeland, J. (1998). Mind embodied and embedded. In J. Haugeland (ed.), *Having Thought*. Cambridge, MA: Harvard University Press.

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

Stich, S., and Warfield, T. (eds.) (1994). *Mental Representation: A Reader*. Cambridge, MA: Blackwell.

van Gelder, T. and Port, R. (1995). It's about time: An overview of the dynamical approach to cognition. In R. Port and T. van Gelder (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

5 Coupling, Constitution, and the Cognitive Kind: A Reply to Adams and Aizawa

Andy Clark

1 Introduction: Crossed Wires

Adams and Aizawa, in a series of recent and forthcoming essays (2001, 2009, this volume) seek to refute, or perhaps merely to terminally embarrass, the friends of the extended mind. One such essay begins with the following illustration:

Question: Why did the pencil think that 2 + 2 = 4?
Clark's answer: Because it was coupled to the mathematician. (Adams and Aizawa, this volume, p. 67)

"That," the authors continue, "about sums up what is wrong with Clark's extended mind hypothesis." The example of the pencil, they suggest, is just an especially egregious version of a fallacy said to pervade the literature on the extended mind. This fallacy, which they usefully dub the "coupling-constitution fallacy," is attributed,[1] in varying degrees and manners, to Van Gelder and Port (1995), Clark and Chalmers (1998), Haugeland (1998), Dennett (2000), Clark (2001), Gibbs (2001), and Wilson (2004). The fallacy, of course, is to move from the causal coupling of some object or process to some cognitive agent, to the conclusion that the object or process is part of the cognitive agent, or part of the agent's cognitive processing (see, e.g., Adams and Aizawa, this volume, p. 68). Proponents of the extended mind and related theses, Adams and Aizawa repeatedly assert, are prone to this fallacy in part because they either ignore or fail to properly appreciate the importance of "the mark of the cognitive," that is, the importance of an account of "what makes something a cognitive agent" (ibid., p. 68). The positive part of Adams and Aizawa's critique then emerges as a combination of the assertion that this "mark of the cognitive" involves the idea that "cognition is constituted by certain sorts of causal process that involve non-derived contents" (ibid.) with the claim that these processes look to be

characterized by psychological laws that turn out to apply to many internal goings-on but that do not currently apply (as a matter of contingent empirical fact) to any processes that take place in nonbiological tools and artifacts.

In what follows, I show why these arguments display nothing so much as mutual failures of communication: crossed wires concealing a couple of real, but rather more subterranean, disagreements. In particular, I show why the negative considerations advanced by Adams and Aizawa fail to successfully engage the argument for the extended mind, and why their more radical positive story, unless supplemented by implausible additional claims, does nothing to undermine the conclusion that minds like ours can (without the need for any radically new techniques, technologies, or interventions) extend into the world.

Before embarking on this, a word about the intended force of the argument. Adams and Aizawa make much of their concession (see, e.g., Adams and Aizawa 2009) that mental extension is *possible*, just not, they claim, actual. Theirs, they insist, is a "contingent intercranialism" applicable to human agents in the current state of technology. But they seem to imply that our view, if it is to stand in contrast to theirs, must be that such extension is rampant, and that "in ordinary tool use we have instances in which cognitive processes span the cranial boundary and extend into intercranial space" (ibid., p. 79). Whatever the truth of such a claim (of rampant extension), it was not the claim made by Clark and Chalmers (1998, reprinted in this volume). Our claim was that in fairly easily imaginable circumstances—ones that involved no giant leaps of technology or technique—we would be justified in holding that certain mental and cognitive states extended (in a sense to be explained later) into the nonbiological world. This leaves it open whether there are such extensions and (if there are) exactly how widespread they are. But it is far stronger than the mere claim of "logical possibility" that Adams and Aizawa suggest as the alternative to rampant actual extension.

2 The Odd Coupling

Consider the following exchange, loosely modeled on Adams and Aizawa's opening "reductio":

Question: Why did the V4 neuron think that there was a spiral pattern in the stimulus?
Answer: Because it was coupled to the monkey.

Now clearly, there is something wrong here. But the absurdity lies not in the appeal to coupling but in the idea that a V4 neuron (or even a group of V4 neurons, or even a whole parietal lobe . . .) might *itself* be some kind of self-contained locus of thinking. It is crazy to think that a V4 neuron thinks, and (just as Adams and Aizawa imply) it is crazy to think that a pencil might think. Yet the thrust of Adams and Aizawa's rhetoric is, again and again, to draw attention to the evident absence of cognition *in the putative part* as a way of "showing" that coupling (even when properly understood—see below) cannot play the kind of role it plays in the standard arguments for cognitive extension. Thus we read that:

> When Clark *makes an object cognitive* when it is connected to a cognitive agent, he is committing an instance of a "coupling-constitution fallacy. (Adams and Aizawa, this volume, p. 67; my emphasis)

But this talk of an object's being or failing to be "cognitive" seems to me almost unintelligible when applied to some putative *part* of a cognitive agent or of a cognitive system. What would it mean for the neuron *or* the pencil to be, as it were, brute factively "cognitive"? Nor, I think, is this merely an isolated stylistic infelicity on the part of Adams and Aizawa. For the same issue arose many times during personal exchanges[2] concerning the vexed case of Otto and his notebook (the example used, with a great many riders and qualifications, in Clark and Chalmers 1998). And it arises again and again, as we shall later see, in the various parts of their recent challenge to engage the issue of "the mark of the cognitive."

Let us first be clear then about the precise role of the appeal to coupling in the arguments for the extended mind. The appeal to coupling is not intended to make any external object "cognitive" (insofar as this notion is even intelligible). Rather, it is intended to make some object, which in and of itself is not usefully (perhaps not even intelligibly) thought of as *either cognitive or noncognitive*, into a *proper part of some cognitive system*, such as a human agent. It is intended, that is to say, to ensure that the putative part is poised to play the kind of role that *itself* ensures its status as part of the agent's cognitive routines.

Now, it is certainly true (and this, we think, is the important fact to which Adams and Aizawa's argument might successfully draw the reader's attention) that not just any old kind of coupling will achieve even this result. But probably no one in the literature, and certainly not Chalmers and I, ever claimed otherwise. Hence the presence of the conditions of (broadly speaking) "glue and trust" pursued at length in the original essay, and briefly summarized in various other places, including the target essays

by Adams and Aizawa. There is no need to repeat the conditions, even summarily, here, as the present focus is on the overall shape of our argument and on issues concerning coupling and the mark of the cognitive, rather than on these aspects of the original content. But it is worth noting that the bulk of our (Clark and Chalmers 1998) treatment was devoted to the isolation and defense of these very features.

The biggest of the crossed wires in the exchange with Adams and Aizawa, we now believe, lies quite close by. For Adams and Aizawa often fail to fully appreciate that the conditions speak to the question (which we deem intelligible) "when is some physical object or process part of a larger cognitive system?" and not to the much murkier question "when should we say, of some such candidate part, that it is *itself* cognitive?" The only question at issue, then, was what kind of coupling makes *for incorporation into* a single cognitive system rather than simple *use by* a cognitive system.

In outlining an answer, we chose to be guided by a set of intuitions derived from reflection on the ordinary use of talk of non-occurrent, dispositional beliefs. In essence, we took these intuitions and systematically showed that the kind of functional poise (poise to guide various forms of behavior) associated with such dispositional believings might be supported by a nonstandard physical realization in which a notebook (for example) acted as the medium of long-term storage. The right kind of coupling to make the external resource into a part of the cognitive system, we argued, was one that poised the information contained in the notebook for sufficiently easy, reliable, and automatic "use" (deployment would be a better word) in much the same way as is typically (though not always) achieved by biological encoding.

Chalmers and I thus offered an argument (which one may accept or reject: that is, of course, another matter) concerning conditions not of "being cognitive" but for incorporation into a cognitive system. In so doing we were not even close, as far as we can see, to committing any simple coupling-constitution fallacy.

We must be cautious, however, for it is not, strictly speaking, that Adams and Aizawa fail to see that the real issue concerns cognitive incorporation. Indeed, they are well aware that the conclusion we were aiming for is that the object or process be part of the agent's cognitive apparatus (see, e.g., Adams and Aizawa, this volume, p. 68). The misunderstanding is more complex, and ultimately more interesting, than that. Adams and Aizawa seem to think that some objects or processes, *in virtue of their own nature* (see section 3 below) are, as we shall now put it, *candidate parts* (for inclusion in a cognitive process), whereas other objects or processes, still in

virtue of their own nature, are not. This, I think, must be the way to give sense to that otherwise baffling question "is some *X* cognitive?" when asked of some putative part. This then is the link between the skirmish concerning a putative coupling-constitution fallacy and the subsequent positive story concerning the "mark of the cognitive." Thus the authors ask:

> if the fact that an object or process *X* is coupled to a cognitive agent does not entail that *X* is a part of the cognitive agent's cognitive apparatus, what does? *The nature of X, of course*. One needs a theory of what makes a process a cognitive process. . . . One needs a theory of the "mark of the cognitive." (Adams and Aizawa, this volume, p. 68, my emphasis)

It is to this (vexed and vexing) issue that we now turn.

3 On Your Marks . . .

So wait a moment: maybe that V4 neuron *is*, in some intelligible sense, cognitive? Maybe it is cognitive in the sense (identified above) of being, *in virtue of its own nature*, at least a *candidate* for becoming a proper part of a genuinely cognitive process. Such, we are at least tempted to think, must be the underlying belief driving much of Adams and Aizawa's otherwise mystifying critique. This slightly puzzling thought thus brings us to the (marginally) more positive part of their discussion, namely their appeal to the "mark of the cognitive."

Notice first that this way of displaying the debate, if correct, already suggests a major concession to the role of coupling. For assume we find some such acceptable (in virtue of its own nature) candidate part. Then what settles the question of whether that part belongs to this cognitive system, or to that one, or (currently) to no cognitive system at all? It is hard to see just what, apart from appeal to some kind of coupling, at some time in the causal-historical chain, could motivate an answer to this subsequent question.

But let's now stick, as Adams and Aizawa insist we should, to the topic of the "mark of the cognitive," and hence to the question (as we see it) of cognitive candidacy rather than actual cognitive incorporation. What could it be that, as they put it, "makes a process a cognitive process" (this volume, p. 68)? The question is nontrivial and has, as Adams and Aizawa somewhat reluctantly admit, no well-established answer within cognitive science or philosophy of mind. But they happily tie their colors to what they depict as "a rather orthodox theory of the nature of the cognitive"

(Adams and Aizawa 2001, p. 52). According to this theory (ibid., p. 53), "cognition involves particular kinds of processes involving non-derived representations." This is the line also pursued in Adams and Aizawa (this volume, 2009). It comprises two distinct elements, just as presented in the quote—namely, an appeal to nonderived content and an appeal to "particular kinds of process."

Despite its prominence in their account, Adams and Aizawa really tell us very little about what nonderived content is. We learn that it is content that is in some sense intrinsic (Adams and Aizawa 2001, p. 48). We learn that this is to be contrasted with, for example, the way a public language symbol gets its content by "conventional association" (ibid.). We are told, in the same place, that Dretske, Fodor, Millikan, and others are (sometimes) in search of an adequate theory of such content, and that the combination of a language of thought with some kind of causal-historical account is a hot contender for such an account. Toward the end of all this, however, the authors make a concession which, I elsewhere argue (chapter 3 of this volume), takes much of the sting out of the tail of the appeal to nonderived content, however (if at all) that elusive concept is to be unpacked. This is the concession that

> Having argued that, in general, there must be non-derived content in cognitive processes, it must be admitted that it is unclear to what extent every cognitive state of each cognitive process must involve non-derived content. (Adams and Aizawa 2001, p. 50)

As I understand it, this concession allows that an external resource, none of whose states or processes or stored representations are themselves intrinsically contentful (assuming we are able to make sense of that notion in some way) might nonetheless be a proper part of some cognitive process. Otto's notebook, to take the obvious example, might be just such a resource, since it is full of inscriptions written in (let's assume) English. Yet Otto's notebook, in the light of this concession, might still figure as part of the supervenience base for some of Otto's dispositional beliefs even while failing itself to be a repository of states with intrinsic content.

Of course, we do not *have* to think of Otto's notebook this way. A more radical response would be to argue that what makes *any* symbol or representation (internal or external) mean what it does is just something about its behavior-supporting role (and maybe its causal history) within some larger system. We might then hold that when we understand enough about that role (and, perhaps, history) we will see that the encodings in Otto's notebook are in fact on a par with those in his biological memory.

In other words, just because the symbols in the notebook happen to look like words of English and require some degree of interpretative activity when retrieved and used, that need not rule out the possibility that they have also come to satisfy the demands on being, given their role within the larger system, among the physical vehicles of intrinsic content.

Nonetheless, there is something quite compelling, I want to agree, about the idea that there is something conventional about the notebook encodings and even about the thought that some parts of any genuinely cognitive system need to trade in representations that are not thus conventional. To accept this, however, is not to give up on the extended mind unless one also accepts (what seems to be an independent and far less plausible assertion) that *no proper part of a properly cognitive system can afford, at any time, to trade solely in conventional representations*. It was this additional claim that, I thought, was being rejected (and, I felt, quite rightly so) in the above quoted passage from Adams and Aizawa.

It seems, however, that I was wrong, and that Adams and Aizawa do in fact endorse something like this additional claim. Thus (this volume, p. 70) the authors accuse me of not seriously attempting to understand the point of their actual concession, and hence of (incorrectly) taking it as rendering the appeal to nonderived content argumentatively vacuous, at least in the case of the debate concerning extended cognition.

So what went wrong? The original concession was followed by an example to which I paid insufficient attention. The example involved possible nonrepresentational elements in a language-of-thought encoding, such as punctuation marks and parentheses (see Adams and Aizawa 2001, p. 50). Such potential elements, they concede, need not count as "intrinsic representations" or even as content-bearing, yet they would still be proper parts of a properly cognitive process. I confess that I simply did not (and still do not) understand this suggestion regarding a language-of-thought encoding (it is repeated in this volume, p. 69, without appearing to me to be any clearer). Nonetheless, it is now clear that whatever it may mean, it was not intended to concede the possibility (given only the considerations concerning intrinsic content) of Otto's notebook counting in the same way. For the authors now clarify their original claim thus:

> Clearly, we mean that if you have a process that involves no intrinsic content, then the [intrinsic content] condition rules that the process is noncognitive. (Adams and Aizawa, this volume, p. 70)

As I now understand it, their position regarding the role of intrinsic content is this: there may be a process that is a genuinely cognitive process

that has as proper parts some goings-on (such as, presumably, the tokening of the punctuation mark in the LOT, puzzling as this still sounds to me) that themselves do not themselves involve intrinsic, nonderived contents (presumably because those parts-of-the-part do not involve contents at all). But such a process (the part, not the part-of-a-part!) must still involve at least *some* intrinsic content on pain of failing to be genuinely "cognitive." And Otto's notebook (I presume they must then wish to assert) fails even this very slightly weakened test, as here (they think) we have a process that involves *no intrinsic content at all*.

But in what sense do we, in the case of Otto's notebook, confront a *process* that involves *no intrinsic content at all*? It helps to be careful about timing here. The time at which the notebook looks most clearly to be part of some real *process* is during the retrieval and use phase, and at that point in time, there are clearly plenty of states in play, in the larger notebook-including system, that count as intrinsically contentful, even on the Adams and Aizawa model. At run time, the process is not one that trades solely in representations whose contents are derived or conventionally determined.

What about at other times? Well, at such other times the claim is just that the notebook is part of the supervenience base for some of Otto's dispositional beliefs. What demands does this make on process? We can at least say this: the very notion of a dispositional belief already makes implicit reference to what would happen in possible run-time situations. So here there is implicit reference to everything that those run-time processes would involve. The poise of the encodings in the notebook is such that, in the appropriate whole-system run-time circumstances, those encodings participate in extended processes that involve (let's assume) states with intrinsic contents.

But suppose, Adams and Aizawa may insist, we put all that run-time process talk aside and look solely at the (putative) part itself. Surely here we find a resource all of whose contentful states are derived, and doesn't that contravene the requirement concerning intrinsic content? In Clark 2003 and 2005b, I offered a thought experiment meant to show that Adams and Aizawa's requirement, as applied to some storage resource considered out of the context of its run-time role in a larger system, was too strong and ought to be rejected. The thought experiment concerned beings ("Martians") endowed with an extra biological routine that allowed them to store *bit-mapped images* of important chunks of visually encountered text. Later on, at will, they could access (and then interpret) this stored text. Surely, I argued, we would have no hesitation in embracing that kind of bit-mapped storage, even prior to an act of retrieval, as part and parcel of the Martian cognitive equipment. But what is stored is just a bit-mapped

image of a fully conventional form of external representation. If we accept the Martian memory into the cognitive fold, surely only skin-and-skull-based prejudice stops us extending the same courtesy to Otto.

Despite spending significant time on what I presented as a weaker and more complex example (the one involving reasoning with imagined Venn diagrams/ Euler circles[3]), Adams and Aizawa do not comment on this case. Yet it raises, I still believe, exactly the right issues. Even if we demand the involvement, in any cognitive process, of at least some items that bear their contents intrinsically, it is quite unclear how we should distribute this requirement across time and space. The Martian encodings are poised, here and now, to participate in processes that invoke intrinsic contents. So are those in Otto's notebook. Since it is arguably poise that matters, at least where dispositional believing is concerned, it seems that any reasonably plausible form of the requirement involving intrinsic content is met.

The notebook, I am happy to concede, is not, considered all on its own (and as far as we understand this notion at all) "intrinsically cognitive." But it *is* a resource whose encodings, at appropriate run-time moments, inform Otto's behavior in the way characteristic (we claimed) of dispositional beliefs. And this, we claim, is all that matters. Perhaps it is indeed essential that any truly cognitive *activity* (and hence any genuinely cognitive *agent*) draw on at least some states with intrinsic content. But we have been given no reason at all to accept the further (and crucial) claim that *no proper part* of such a properly cognitive system, considered now in splendid isolation from those crucial run-time wholes in which it participates, can afford to contain only representations lacking intrinsic content.

Indeed, I see no reason why we should accept (or even be tempted by) such a further condition. In general, for some *X* to be part of the supervenience base of some *Y*, where that *Y* must (to count as a *Y* at all, let's assume) exhibit some property *Z*, there is no requirement *that Z be in addition a property of the putative part X*. Thus suppose it were essential, for any system to count as properly cognitive, that the system be capable of conscious awareness. We would not want to insist (indeed, we would be crazy to insist) that every proper part of that system be capable of such awareness. We would not even insist (to draw even closer to the case in hand) that every proper part *of the subsystems that support conscious awareness* need be such as to exhibit such awareness when considered in isolation. Or suppose that we think that any genuinely moral agent must be able to reason about the good of others. Still, we should not think that every proper part of that agent (not even every proper part essential to the agent's moral reasoning) must be capable of so doing. Just so, from the