Cognitive Spread: Under What Conditions Does the Mind Extend Beyond the Body?

Zed Adams and Chauncey Maher

Abstract: The extended mind hypothesis (EMH) is the claim that the mind can and does extend beyond the human body. Adams and Aizawa (A&A) contend that arguments for EMH commit a 'coupling constitution fallacy'. We deny that the master argument for EMH commits such a fallacy. But we think that there is an important question lurking behind A&A's allegation: under what conditions is cognition spread across a tightly coupled system? Building on some suggestions from Haugeland, we contend that the system must exhibit a distinctive sort of semantic activity, semantic activity that the system as a whole takes responsibility for.

'An individual's *being* responsible is its *taking over* responsibility for its *whole* self'.
—John Haugeland (2000: 65)

1. Introduction

Many of us have a peculiarly intimate relationship with our iPhones. We bring them everywhere, all the time, and they play an essential role in how we navigate our environment, communicate with others, remember our experiences, and plan for the future.

This and examples like it have been taken by many philosophers and cognitive scientists to illustrate the extended mind hypothesis (hereafter EMH).[1] According to EMH, the human mind is a tightly coupled system that extends beyond the human body to include external, worldly things, such as our iPhones.

Many people are unmoved by such examples.[2] Fred Adams and Ken Aizawa (hereafter A&A) allege that not all coupling is constitution; to think otherwise is to commit the 'coupling-constitution fallacy'.[3]

In this essay, we explain why the master argument for EMH does not commit such a fallacy.[4] But we admit that there is an important question lurking behind A&A's allegation: under what conditions is cognition spread across a tightly coupled system? Building on some suggestions from John Haugeland, one of the originators of EMH, we contend that the system must exhibit a distinctive

sort of 'semantic activity', semantic activity that the system as a whole takes responsibility for.[5]

In §2.1, we sketch the master argument for EMH. In §2.2, we present and undermine the allegation that the master argument commits the 'coupling-constitution fallacy', explaining what might have led A&A astray. In §3, we observe that, nevertheless, one should like to know when cognition is spread across a tightly coupled system. In §§3.1-3.3, we sketch a controversial answer, explaining that the system must engage in a distinctive sort of self-criticism.

2. EMH and the Coupling-Constitution Fallacy

2.1. The Argument for the Extended Mind Hypothesis

In this section, we summarize the master argument for EMH, as it originated with Haugeland (1995) and was developed by Andy Clark and David Chalmers (1998).[6]

When arguing for EMH, the main burden is to show that external items are genuine constituents of a cognitive process, not merely causal influences on it. The distinction between constituents and causal influences on a process is difficult to formulate in general terms, but is easy to illustrate with examples. Consider the case of playing a game of chess: the chessboard and pieces are constituents in this process; the pizza eaten while playing is merely a causal influence on it (e.g., if it provides the energy necessary for moving the pieces). There is a relationship of dependence in both cases—between a process and its constituents and a process and its causal influences—but it is a different sort of dependence in each case. The causal influences on a process play a role in bringing it about; the constituents make it what it is.[7]

Haugeland focuses on a case of navigation. How does a normal person find her way from Berkeley to San Jose, California? Here is his wonderfully terse account of how *he* finds the way: 'I pick the right road (Interstate 880 south), stay on it, and get off at the end' (1995: 234). Haugeland claims that the road should count as part of the cognitive system responsible for finding the way. By this, he does not mean that the road is a mere cognitive aid—something external to his mind but which his mind uses to help find the way. Rather, he means that he and the road form a single cognitive system.

To make his case, drawing upon principles of systems analysis first articulated by Herbert Simon, Haugeland distinguishes between *high*- and *low-bandwidth interaction*. He points out that spatial contiguity is not a good way to identify the components of a system. For instance, in many universities, the offices of the philosophy department are spread throughout the campus, across

different buildings. Offices need not be contiguous in order to be a component of a university, such as a department. A much better way to identify the components of a university is to look at the nature of the interaction between offices, regardless of their contiguity. Certain offices interact in an intense, tightly-coupled manner (through meetings, email, etc.), thereby functioning as part of a relatively independent and self-contained component of the university as a whole. Other offices interact less intensely: they simplify the products of their work and exchange these simplified reports with each other. High-bandwidth interaction is intra-component interaction; it is what goes on *within* a part of a system. Low-bandwidth interaction is inter-component interaction; it is what goes on *between* parts of a system. A job hire meeting, in which the members of a department work together to decide who to hire, is an example of the sort of intense, tightly coupled interaction that is characteristic of high-bandwidth interaction; reporting to the dean's office about the results of such a meeting is the kind of simplified interaction that is characteristic of low-bandwidth interaction.

Haugeland argues that when he finds his way to San Jose, he is in a high-bandwidth interaction with the road. In this respect, the road functions just as an inner mental representation of the way to San Jose would function.[8] On those grounds, Haugeland claims that the road itself should count as part of the cognitive system. In his words:

> [M]uch as an internal map or program, learned and stored in memory, would … have to be deemed *part of* an intelligent system that used it to get to San Jose, so I suggest the *road* should be considered *integral* to my ability. (1995: 234)

Thus, Haugeland thinks that the mind extends well beyond the central nervous system.[9]

Clark and Chalmers (C&C) make a similar argument. They focus on a case of memory, asking us to imagine Otto, an Alzheimer's patient who remembers where the Museum of Modern Art is by writing down its address in a notebook. There are four crucial features of Otto's interaction with his notebook: (i) Otto always carries the notebook with him; (ii) the notebook itself is easy to access; (iii) Otto automatically endorses whatever he reads in it; (iv) Otto only writes things in it that he has explicitly endorsed in the past. Taken together, these four features imply that this is an example of high-bandwidth interaction. Given this, C&C argue that the contents of the notebook—e.g., MOMA's address—are functionally equivalent to stored mental representations (i.e., memories). If memories are part of Otto's mind, then so too are the contents of the notebook. C&C offer the following principle by which they arrive at this conclusion:

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. (1998: 222)

Since the contents of the notebook are part of Otto's mind, his mind extends beyond his body.

It is important to see that Haugeland's and C&C's arguments share a structure. They argue that the mind extends because there are external items (a road, a notebook) that are functionally equivalent to admitted parts of the mind (a map, a memory). And these external items are functionally equivalent to parts of the mind because the external items have the same high-bandwidth interactions that those parts of the mind have to other parts of the mind.

Schematically, here is the master argument for EMH:

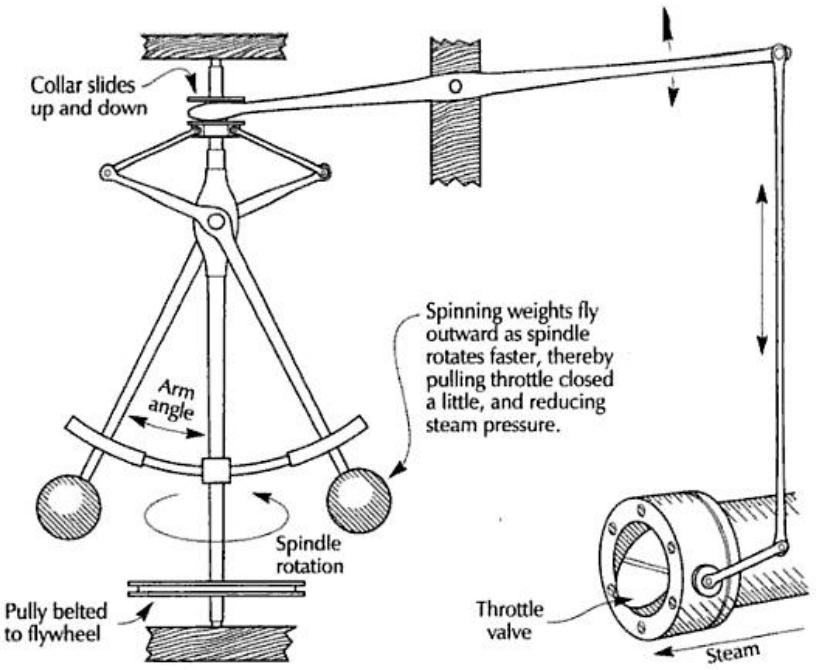
1. Y is part of a cognitive system Z.
2. X (an external item) has the same high-bandwidth interaction with other parts of Z that Y has.
3. So, X is functionally equivalent to Y.
4. So, X is part of Z.

2.2. The Coupling-Constitution Fallacy

In a series of publications, Fred Adams and Ken Aizawa have insisted that not all coupling is constitution. They allege that proponents of EMH tend to commit a 'coupling-constitution fallacy', the mistake of thinking that because something is coupled to a mind, it must be part of that mind. Here is their schematic account of this fallacy:

The coupling-constitution fallacy:
X is coupled to Y.
So, X is a part of Y.[10]

This pattern of inference is obviously fallacious, whether the topic is minds or anything else. Counter-examples abound. A&A point out that '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' (2010: 68) but there are countless such counter-examples. Consider the classic example of a tightly coupled system, Watt's steam engine governor.[11]



12

In this tightly coupled system, the spinning weights (on the left) are *tightly coupled* to the throttle valve (on the right), but that does not imply that the spinning weights are *part* of the throttle valve.

The ease with which one can come up with such counter-examples suggests that it is not a plausible characterization of anything anyone actually thinks.[13]

Indeed, the coupling-constitution fallacy simply is not an accurate representation of the master argument for EMH. Specifically, it does not accurately represent the role coupling or high-bandwidth interaction plays in that argument. (In what follows, we will use 'coupling' and 'high-bandwidth interaction' and their cognates interchangeably.)

First, nowhere does the master argument claim or imply that if X is coupled to Y, then X is part of Y. Haugeland, for instance, does not claim that since the road to San Jose is tightly coupled to him it becomes part of him. Rather, he claims that the he and the road together make up a coupled cognitive system. Schematically, here is Haugeland's actual claim:

> X is coupled to Y.
> So, X and Y make up Z.

The idea is that when two things are tightly coupled, they constitute some third thing, of which they are both parts. A&A do not seem to recognize that.

Second, while the master argument for EMH certainly appeals to coupling, it is not merely an argument from coupling. As we saw, Haugeland argues that the

road is functionally equivalent to an 'internal map or program' because it has the same kind of high-bandwidth interaction with him that an 'internal map or program' would have. Because the road is functionally equivalent to an admitted part of a mind, it is part of a mind. Thus, coupling is only one step in the argument for EMH. Recall our schematization of the master argument:

1. Y is part of a cognitive system Z.
2. X (an external item) has the same high-bandwidth interactions with the rest of Z that Y has.
3. So, X is functionally equivalent to Y.
4. So, X is part of Z.[14]

Thus, A&A misconstrue the place of coupling in the master argument for EMH. Having gotten clear about its place in the argument, it is not obvious that the argument as a whole is fallacious. Indeed, the argument looks plausible.

Two questions naturally arise at this point. First, what explains A&A's misconstrual of the argument for EMH? Second, is there anything left of A&A's worries?

A preliminary explanation for A&A's misconstrual is that they seem to conflate coupling with (mere) causal interaction. This is evident in the examples of 'coupling' that they themselves introduce to the discussion. For instance, they suggest that Clark holds that a pencil thinks simply because it is 'coupled' to a mathematician who uses it to write the equation '2+2=4'.[15] A&A contend that in this case the mind obviously does not extend, so Clark is wrong about what it takes for the mind to extend.

However, Clark and other defenders of EMH would (or should) simply agree that in this case the mind does not extend. It does not extend because the pencil (and paper) play no constitutive role in calculating the answer to this equation. More importantly, and more precisely, cognition does not extend to the pencil because it is not coupled to the relevant parts of the mathematician in her act of calculating. Although the mathematician certainly uses the pencil to write the equation, there is no good reason to think that what she writes is the result of this interaction. Rather, the mathematician first computes the answer to the equation in her head and only then records this answer on the paper. The pencil and paper are simply instruments for recording the result of a process completed elsewhere. *Writing* '2+2' is not part of *computing* the answer to the equation. By contrast, consider a case of multi-digit multiplication: '49885320 x 12534959 = 625310440901880'. When one uses a pencil, paper, and a long-multiplication algorithm to compute the answer to that equation, writing plays a constitutive role in the process of figuring out the answer, since the position of the numerals that one writes on the paper are *part* of how one arrives at the solution to the equation.[16] A&A's example of writing '2+2=4' is fundamentally unlike this, for

the position of the numerals written on the paper does not play a role in computing the answer to the equation at all; we can explain what the mathematician is thinking and why she is thinking it without making any reference to the pencil and paper whatsoever. A&A's example is thus more like the way in which written mathematical notation was used in the ancient world, in that '[a]ll sizable calculations in the ancient world were performed with the aid of some kind of abacus; a written number representation was needed for record purposes only' (Hollingdale and Tootill, 1965: 21). Thus, A&A seem to treat a case of (mere) causal interaction as a case of coupling.

There is a deeper explanation for A&A's misconstrual, however. To see this, it will help to talk of embedded and unembedded bodies. Focus on Haugeland's trip to San Jose. In the actual case, his body is embedded, for it is tightly coupled with the road. To be embedded with something is to be coupled to it. It is possible, however, to imagine a counter-factual case in which he has an inner mental representation of the route to San Jose. Imagine, for instance, that Haugeland is able to find the way to San Jose by memorizing the number of steps and turns that he must take in order to get there from a certain starting point. In this imaginary case, his body would be unembedded, for although his body would be *causally interacting* with the road, it would not be *coupled* with it. Strictly speaking, in this imaginary case, the road does not play any cognitive role whatsoever in Haugeland's finding his way to San Jose.

Now, advocates of EMH take the actual case to be fundamentally different from the imaginary case. A&A, however, would see the cases as fundamentally similar. Because they conflate coupling with mere causal interaction, they would emphasize a superficial similarity between the embedded and unembedded cases: in both, Haugeland's body causally interacts with the road. For A&A, this similarity suggests that whatever cognitive capacities explain his behavior in the imaginary case, they must also explain it in the actual case. In other words, they would assume that since in the imaginary case his body functions as a self-contained cognitive system, it must function similarly in the actual case. But this assumption is a mistake, and the superficial similarity it rests upon is deeply misleading. Although Haugeland's body causally interacts with the road in both cases, the functional role of his body is fundamentally different in them: in the imaginary case, his body is unembedded; in the actual case, his body is embedded.

In short, by conflating causal interaction and coupling, A&A appear to hold that the cognitive capacities of unembedded bodies are just the same as the cognitive capacities of embedded bodies. But this is precisely the assumption that the master argument for EMH is designed to put in doubt. A&A's mistake is to assume that the parts and wholes of cognitive systems are fixed by how they

sometimes function. That is, they assume that because Haugeland himself *sometimes* functions as a self-contained cognitive system, he must *always* function as such.[17] The case of the road to San Jose is designed as a counter-example to precisely that inference. A&A, therefore, are themselves guilty of a fallacy, that of assuming that because something *sometimes* functions in a certain way, it must *always* function in that way. We might call it the functional fixedness fallacy. When Haugeland is finding the way to San Jose, he is only part of a larger cognitive system, one that contains the road as well.

3. Giving a Damn

Having diagnosed the sources of A&A's misconstrual of the argument for EMH, a second question remains: is there anything left of A&A's worries? That is, although A&A are wrong to accuse advocates of EMH of committing the coupling-constitution fallacy, is it nonetheless possible to extract a more legitimate worry from their writings about EMH? We think it is.

When does coupling involve (what we will call) *cognitive spread*? Consider the case of Nat. One day, Nat decides to go bungee jumping. When Nat jumps off the platform, his body and the bungee cord form a tightly coupled system, since a proper understanding of the behavior of his body must make essential reference to the behavior of the cord, and vice versa. But none of Nat's cognitive processes are spread throughout the bungee cord. The question raised by A&A is what, if anything, is the difference between Nat's case and the cases of John (Haugeland) and Otto? That is, when does coupling X (something that sometimes functions as a self-contained cognitive system) and Y (something that, on its own, never functions as a self-contained cognitive system) make Z (a new cognitive system, one which is not merely another way of referring to X)? In short, under what conditions is cognition spread across a coupled system? What, if anything, distinguishes the coupling of Nat+cord from the coupling of Otto+notebook?

3.1. What makes something a *cognitive* system?

As Nat's bungee cord case clearly illustrates, coupling need not imply cognitive spread. The most obvious way to address this worry is to identify what makes something a cognitive system in the first place. Following A&A, this is the strategy we propose to adopt:[18]

> The bounds of cognition must be determined by finding the mark of the cognitive, then seeing what sorts of processes in the world have the mark. (2001: 46)

There are two aspects to this strategy: first, to identify a plausible mark of the mental; second, to see if this mark is spread throughout the tightly coupled systems discussed by Haugeland and C&C. Unlike A&A, we think that the most plausible such mark is, in fact, spread throughout these systems.

We agree with A&A that the most plausible mark of the mental is intentionality: the property of being about or directed toward something. Taking intentionality to be the mark of the mental arguably originates with Franz Brentano:

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction towards an object (which is not to be understood here as meaning a thing), or immanent objectivity. (Brentano, 1874/1995: 88–89)

Taking intentionality to be the mark of the mental is contested, of course;[19] it is enough for our purposes to note that it is the mark of the mental most often invoked as a way of *criticizing* EMH. That said, not just any sort of intentionality will do, if only because not all sorts of intentionality are sufficient for mindedness. Consider, for example, cases in which we explicitly stipulate that something has intentionality: for instance, we can stipulate that the shape '⊕' refers to the Empire State Building. Having done this, '⊕' now has intentionality, but that hardly suffices to show that it is minded. In order to identify the sort of intentionality that is sufficient for mindedness, it will help to draw a distinction between intentionality that is merely derivative—for it exists only in virtue of something else bestowing intentionality on it (as we have just done with the shape '⊕')—and intentionality that is not derivative. As A&A and other critics of EMH note, the sort of intentionality that is needed to show that something is minded is this second sort of non-derivative, or original intentionality.[20]

Merely drawing a distinction between original and derivative intentionality and claiming that original intentionality is the mark of the mental is hardly sufficient for settling the debate between EMH and its critics, however. What is needed is a non-question begging characterization of what it is for a system to possess original intentionality, one that does not simply assume that Otto's notebook—or anything else—either does or does not have original intentionality. What is it, for instance, for Otto's CNS to possess original intentionality? What differentiates the intentionality possessed by Otto's CNS from the intentionality possessed by '⊕'? We think that it is necessary to answer

this question in order to respond to the legitimate worry raised by A&A, the question of when, exactly, cognition is spread across a tightly coupled system.

3.2. Semantic Activity

What is the mark of original intentionality? It will not do simply to point to examples of systems that possess original intentionality, such as the CNS, if only because the debate over EMH is about whether things *other* than the CNS can possess original intentionality.[21] We think a good starting point for identifying the mark of original intentionality lies in a characteristic pattern of activity that Haugeland calls 'semantic activity'. Haugeland introduces the idea of semantic activity by reflecting upon the following difference between printed words that possess mere derivative intentionality (such as the '⊕') and brain states in the CNS that possess original intentionality. He writes:

> The most conspicuous difference between book symbols (printed words) and brain symbols (thoughts) is that book symbols just sit there, doing nothing, while thoughts are constantly interacting and changing. But that by itself can't be the crucial difference, since book-like symbols can also be interactive and variable. Imagine printing individual word tokens on thousands of tiny floating magnets or on the backs of busy ants. The resulting physical symbols would interact vigorously and form various amusing combinations; but their meaning would still be just as derivative (from us) as those of any lines inked quietly on a page.
>
> Nevertheless, this may be on the right track. Consider writing the following premises on a piece of paper:
>
> All frogs eat insects.
> Fido is a frog.
>
> Nothing much happens. More revealing: if we set them out on magnets or ants, then plenty of activity might ensue, but most likely nothing interesting about frogs or insects. In sum, the behavior of such symbols (independent of our manipulations) is quite unrelated to what they mean. By contrast, if we got those premises into somebody's head (say, in the form of beliefs), then a new complex symbol would very likely appear—and not just any new symbol, but specifically a valid conclusion about Fido and insect eating:
>
> Fido eats insects.
>
> So the important difference is not merely that thoughts are active, but that their activity is directly related to their meanings; in general, the changes and interactions among thought symbols are semantically appropriate. Accordingly, I say that thought symbols are *semantically active*, in

> contradistinction to printed or magnetic symbols, which are semantically inert (even if they're 'active' in other ways that are semantically irrelevant). (1985: 120-1)

Haugeland's proposal is that systems that have original intentionality exhibit semantic activity; systems that have mere derivative intentionality exhibit intentionality, but are not semantically active. The proposal is attractive because it captures the intuition that things with *original* intentionality are 'active' in a way that things with *derivative* intentionality are not (and it does so without simply assuming that brain states have original intentionality because they are brain states).

Alas, there is a problem with this proposal. Calculators are semantically active and yet it is implausible to think that they have original intentionality. Thus, this proposal seems in need of qualification. Not just any semantic activity is sufficient for original intentionality.

What more is needed for original intentionality?

3.3. Giving a Damn

In this section, we propose an account of the sort of semantic activity that suffices for original intentionality. Our proposal is controversial. We fully expect that philosophers who find EMH counter-intuitive will find our proposal just as counterintuitive. That said, our proposal is not without precedent. Broadly speaking, it can be located in the tradition of thinkers who claim that a distinctive sort of behavioral flexibility is required for genuine thought.[22] We propose that a system must give a damn about its semantic activity for that activity to exhibit original intentionality.[23] More precisely—drawing on another of Haugeland's ideas—we argue that a system must take responsibility for the proper functioning of its semantic activity for that semantic activity to exhibit original intentionality.

This section has two parts. First, we explain what sort of responsibility is required for original intentionality. Second, we argue that this sort of responsibility is exhibited by the Otto+notebook system but not by the Nat+cord system. We will thus provide an answer to the question that opened §3: what, exactly, is the difference between these two systems, such that cognition spreads in one but does not spread in the other?

Our proposal is that a system must take responsibility for the proper functioning of its semantic activity for that semantic activity to exhibit original intentionality. By this, we mean that it must be critical of its semantic activity. There are three levels to such self-criticism. We will talk about each in turn.[24]

The first level of self-criticism concerns the active manipulation of semantic items (e.g. beliefs, desires, etc.). A system with original intentionality

must abide by norms for producing, manipulating, and removing these semantic items. For instance, cognitive systems should reject beliefs that it discovers to be false. Suppose Otto believes there is cereal in the cupboard, but opens it and sees that there is not any there; all else being equal, he should now reject his belief that there is cereal in the cupboard.[25] The 'should' here is not merely alethic but normative. That is, it does not concern merely what a system does or is disposed to do, but what that system ought to do—if it is to function properly at all. For instance, Otto might not actually give up his belief about the cereal, but he *ought* to give it up on pain of diminished cognitive functioning.

Of course, it is natural to wonder which norms a system should follow. This points to a second level of self-criticism. A system with original intentionality must scrutinize its own norms for modifying the semantic items in that system. So, not only must such a system modify semantic items in accordance with its norms, it must also modify, where appropriate, those norms themselves. The idea is that a system should reject norms that do not work, and adopt norms that do. For instance, Otto might realize that beliefs he forms while drunk tend to be unreliable; he might then adopt a norm of questioning or simply rejecting beliefs identified as such. (Such a realization or such an adoption need not be conscious.)

What shows whether a norm works? As with the first level of self-criticism, the choice of norms is, by-and-large, determined by the role they play in making possible skillful coping with the world. Intuitively, we can say that a norm works if compliance with it tends to result in semantic items that allow for such skillful coping, e.g. true beliefs, good desires, etc.[26] In that respect, a system with original intentionality aims for its success at coping with worldly affairs to be non-accidental. Such success is what is at stake in second-order self-criticism. That is, a system with original intentionality ought to scrutinize its norms *because* it aims at non-accidental success. Norms should be revised so as to satisfy that goal. That points us to a third level of self-criticism.

At bottom, a system with original intentionality must care (i.e., give a damn) about getting things right.[27] Such care is apparent in the first two levels of self-criticism, but it is clearest in the third level of self-criticism. A system with original intentionality must be prepared to give up its whole way of doing things if it realizes that success at the second level was *merely* accidental. That is, it must be prepared to recognize that despite its best efforts to adjust its norms in the light of errors, it is unable to come up with norms that really work, norms that can be *relied* upon to get things right. To give up one's whole way of doing things involves more than rejecting individual norms governing one's semantic activity. Rather, it involves abandoning those norms that are definitive of the system's way of engaging in semantic activity as such.[28] The rejection of those norms (and the

potential adoption of others) is so radical as to constitute an end to the system's old way of engaging in cognitive activity. In that sense, it is like death (and potential rebirth). Thus, good examples of this sort of giving up are cases in which a scientist abandons her existing scientific practice altogether, usually in favor of another: e.g., Copernicus's abandonment of Ptolemaic theory. Copernicus was well aware that Ptolemaic theory was capable of making accurate predictions of the future behavior of heavenly objects. But he came to think that the very project of trying to explain the moon's, planets', and sun's behavior in a shared fashion was mistaken, that the success of the Ptolemaic model in this regard was merely accidental and did not represent a reliable system of norms for getting things right. This is a paradigmatic instance of the third level of self-criticism.

In sum, there are three levels of self-criticism in which a system must engage to exhibit original intentionality, all of which are anchored in caring about getting things right:

> First level: produce and reject semantic items in accord with norms for proper production and rejection;
>
> Second level: produce and reject those very norms, where appropriate;
>
> Third level: give up the whole system of norms in favor of some other.

What remains to be shown is how the tightly coupled system of Otto+notebook exhibit all three levels of self-criticism, and thus counts as a system with original intentionality.

Imagine that Otto desires to visit the Museum of Modern Art. He looks in his notebook and sees this:

> MOMA: 11 East 53rd St.

He goes to that address, but sees that the Museum of Modern Art is not there. He should—and presumably will—question the accuracy of what is written in his notebook. He might not immediately erase it; he might put a question mark by it. He might say aloud, 'What's wrong with this address? Is this the right street, but the wrong number? Is this the wrong street? Are both the number and the street wrong?' He looks at his notebook again, and sees this:

> MOMA Phone: (212) 708-9400

He calls the number and asks what the address is. They tell him it is 11 *West* 53rd St. At this point, he erases 'East' and replaces it with 'West'. This illustrates the first level of self-criticism.

Now, imagine that Otto has gotten drunk several times in the past few weeks. Each subsequent morning, he has found information in his notebook that eventually leads him astray. On one occasion, he had the wrong name for a new acquaintance; on another, he listed a Mexican restaurant as recommended when he had been warned against it; on yet another, he listed himself as having plans for a date that he *wanted* to have but that he had not actually planned. Having been repeatedly led astray by such information, he eventually recognizes a pattern. The information he records when drunk is unreliable. He should do something. But what, exactly, should he do? Should he stop drinking? Should he stop recording notes while drunk? Should he be especially attentive when recording notes while drunk? These are the sorts of question that should arise as he contemplates what would be the best way for him to get notes that are more consistently accurate or true. Suppose he decides to be more attentive when recording notes while drunk, but it does not work. He finds that the pattern persists: even when more attentive, the notes are consistently inaccurate. He will then have to try something else. Whatever exact policy he settles on, his attempt to find one that works (or works *better*) illustrates the second level of self-criticism.

Now, imagine that there are more persistent and widespread problems. Suppose that Otto develops Parkinsons's Disease, and with it regular, intense hand tremors. They make it very hard for Otto to write. Not only is it unpleasant for him to do so but what he writes is illegible, even to him. Unlike the notes written while drunk, this problem is not localized, but pervasive, for the tremors occur regularly. Moreover, despite consultations with specialists, the tremors cannot be brought under control. In such a situation, Otto might well abandon the use of the notebook altogether, *even if* he does not have some other, more reliable, method of remembering information. This is an example of the third level of self-criticism, since it involves giving up being a certain sort of person, one who records memories in a notebook.

Now contrast the cases of Otto+notebook and Nat+cord. The difference that makes a difference (for our purposes) is that *only* the notebook exhibits the right sort of semantic activity. Consider the following three differences between Otto's notebook and Nat's bungee cord. First, neither the bungee cord nor anything on it represents anything. The notebook is full of *semantic items*: the writing in it represents names, locations, times, etc. Thus, the notebook is plausibly a site of intentionality—whether it is derivative or original is a separate issue. Second, since the cord does not represent anything, there is certainly no *semantic activity* there. The semantic items in the notebook, however, are semantically active; they change according to their meanings.[29] Now, such activity by itself does not suffice for original intentionality; the activity has to be

of the right sort, which leads us to the third and most important difference. Since there is no semantic activity in the cord, Nat+cord cannot engage in self-criticism of such activity.[30] However, as we have just seen, Otto+notebook does take responsibility for the semantic activity in the notebook. That suggests that the notebook is not just a site of intentionality, but of *original* intentionality. The notebook is not merely an optional device for recording the results of semantic activity that originally takes place somewhere else. Rather, it is where that semantic activity originally takes place. As such, it is a constitutive part of a system that exhibits original intentionality.[31]

It is tempting to retort that Otto+notebook *as a whole* does not take responsibility for the semantic activity in the notebook; Otto's body does all of the work. In a trivial sense, this allegation is true: Otto's body writes in the notebook; the notebook does not write on Otto's body. But, in a more substantial sense, the allegation is either false or irrelevant. The allegation is false if it is supposed to mean that the notebook does not play an essential role in the behavior of the system as a whole. Consider the case of Inga, who has brain-bound memories of MOMA's address. Just as Inga's memories determine where she looks for MOMA, the contents of Otto's notebook determine where he looks for it. Likewise, just as Inga should revise her memory if she does not find MOMA where her memory says it is, so Otto should revise his notebook if he does not find MOMA where his notebook says it is. Neither Inga nor Otto would be able to find the museum if these parts of their cognitive systems malfunctioned. (They are, in this sense, functionally equivalent.) The allegation is irrelevant if it is supposed to mean that Otto's notebook does not, by itself, take responsibility for the proper functioning of the system. By parity of reasoning, if every *part* of our minds had to take responsibility for the proper functioning of the *whole*, then Inga's memories would not count as part of her mind![32]

4. Conclusion

This paper has three goals. First, to offer a perspicuous account of the main argument for EMH and to identify the most prominent objection to it. Second, to undermine this objection, not just by showing that it misconstrues the main argument for EMH but also by diagnosing the source of this misconstrual. Third, to point the way forward for EMH by outlining a view of the nature of cognitive systems that *explains* what makes cognitive spread possible.

If there is one point to take away from this paper, it is that previous discussions of EMH (both by advocates and detractors) have not adequately explored the nature of original intentionality. They have tended to rest content with an intuitive understanding of it as the mark of the mental, treating it as *whatever it is* that our unembedded bodies do when they think about things. But

that idea leaves it utterly opaque under what conditions original intentionality—cognition—might extend beyond our bodies. We have begun to correct that opacity, trying to explain what original intentionality consists in and under what conditions it can extend beyond the human body. Of course, we have offered only a sketch, one that deserves to be extended.[33]

Zed Adams
The New School for Social Research
New York, New York 10003
zedadams@gmail.com

Chauncey Maher
Dickinson College
Carlisle, PA 17013
maherc@dickinson.edu

NOTES

[1] For instance: (Haugeland, 1995), (Clark, 1997), (Clark & Chalmers, 1998) (Clark, 2008), (Hurley, 1998), (Rowlands, 1999), (Noe, 2004), (Wilson R. , 2004), (Shapiro, 2011).

[2] For instance: (Adams & Aizawa, 2001), (Grush, 2003), (Rupert, 2004), and (Fodor, 2009).

[3] For instance: (Adams & Aizawa, 2001), (Adams & Aizawa, 2008) (Adams & Aizawa, 2010a), and (Adams & Aizawa, 2010b).

[4] For instance: (Fisher, 2009), (Walter & Kyselo, 2009), (Wilson R. A., 2010).

[5] Clark may be the most prominent advocate of EMH, but he clearly acknowledges the influence of (Haugeland, 1995) on his view. See (Clark, 2008a: 37-38); (Clark, 2008b: xxvi-xxvii).

[6] Clark credits (Varela, Thompson, & Rosch, 1991) as the earliest argument for EMH. But they argue for *embodied* cognition, not *extended* cognition. The argument for EMH outlined here, in terms of general principles of systems analysis and the functional equivalence of external worldly items and inner mental representations, is first clearly articulated in (Haugeland, 1995). That said, it is a mug's game to try to identify *the* origin of a philosophical idea or argument. As Haugeland himself emphasizes, there is a long history to all of the component parts of (what we are calling) the master argument for EMH. Haugeland merely combines these parts in a new and novel manner.

[7] For a different account of the difference between constituents and causal influences, see (Shapiro, 2011: 159-160).

[8] Haugeland uses 'functional equivalence' at (1995: 213).

[9] Haugeland does not seem interested in the question of *whose* mind the road to San Jose is part of; that is, he appears to think that it is a mistake to ask about the boundary conditions of *individual* minds (e.g., as a way of individuating them). To us, this seems like a sensible conclusion to draw from EMH, though it has not—to our knowledge—been noted by subsequent advocates (or detractors) of EMH. Perhaps Descartes' most lasting legacy will be the persistence of the assumption that your mind is *your* mind—that it is uniquely and solely yours.

[10] In the words of A&A, 'It simply does not follow from the fact that process X is in some way causally connected to a cognitive process that X is thereby part of that cognitive process' (2008: 91). Elsewhere, they write, 'the fact that object or process X is coupled to object or process Y does not entail that X is part of Y' (2010a: 68).

[11] The classic discussion of the significance of the Watt steam engine governor for EMH is (Van Gelder, 1995).

[12] Image by John Haugeland. From (Van Gelder, 2007: 425).

[13] Indeed, as A&A note, the most common response to their discussion of the coupling-constitution fallacy is simply 'to say that no one really commits it' (Adams & Aizawa, 2010b). In fact, examples in which X being tightly coupled to Y implies that X is part of Y are comparatively rare: chains and trains are some of the few examples. We are indebted to Justin Horn for these examples

[14] To be clear: we are not claiming that high-bandwidth interaction is identical to functional equivalence. We are claiming that sameness of high-bandwidth interaction is sufficient for being functionally equivalent to a part of a cognitive system. It is not the only way of achieving functional equivalence. Nor is it necessary for functional equivalence. As such, they are not identical. Thanks to an anonymous referee of this journal for indicating that we should be explicit about this.

[15] (Adams & Aizawa, 2010a: 67). When proponents of EMH say that the mind extends to worldly items like iPhones and notebooks, they are not claiming that those items think. Rather, they are claiming that those items are parts of cognitive processes or systems, larger wholes that think. Just as memories and internal maps themselves do not think, iPhones and notebooks themselves do not think.

[16] Multi-digit multiplication is thus similar to the case of the Watt steam engine governor, in which making sense of the behavior of the flyballs necessarily involves making reference to the behavior of the throttle valve, and vice versa.

[17] It is worth noting that the mistake A&A make here is parallel to the mistake made by (Searle, 1980). Searle assumes that because he himself *sometimes* functions as a self-contained cognitive system (e.g., when he is speaking English), he must *always* function as such (e.g., when he is part of a system capable of speaking Chinese). As a general pattern of inference, this is rather obviously fallacious: a particular brigade may function as a self-contained unit in some battles but not in others. See (Haugeland, 1990) and (Haugeland, 2002b).

[18] For the sake of argument, we follow A&A in adopting the strategy that they propose for setting the debate between advocates of EMH and their critics. But we think this strategy exhibits a troublesome indifference to distinctions between parts and wholes in A&A's discussions of EMH. A better, but more difficult, strategy would be to identity the features that make inner memories (for instance) *parts* of cognitive systems and then look to see if there are any external items (such as the contents of Otto's notebook) that exhibit these same features. The problem with A&A's strategy is that the features that make something a *whole* cognitive system need not be possessed by any (let alone all) of its *parts*.

[19] For a useful summary of debates on the question of whether intentionality is the mark of the mental, see (Crane, 1998).

[20] A&A write, 'A[n] essential condition on the cognitive is that cognitive states must involve intrinsic, non-derived content' (2001: 48). Another prominent critic of EMH, Jerry Fodor, writes, ''Underived' content (to borrow John Searle's term) is the mark of the mental; underived content is what minds and only minds have' (Fodor, 2009). 'Underived' content is not, in fact, Searle's term but Haugeland's. As footnote 6 in (Haugeland, 1990) painstakingly details, in the late 1970s both Haugeland and Searle introduced similar but non-equivalent distinctions between

different types of intentionality. Haugeland's distinction was between original and derivative intentionality, where 'original' simply means 'non-derivative'; Searle's distinction was between intrinsic and observer-relative intentionality, where 'intrinsic' means something much more than 'non-derivative'—it means (something like) 'a causal consequence of the physical/chemical/biological structure of the brain'. It seems clear that Fodor means to be invoking Haugeland's distinction, not Searle's. The confusion is understandable, insofar as in more recent writings Searle has (by-and-large) adopted Haugeland's distinction, thereby abandoning the distinction he himself introduced to the literature.

[21] It is worth noting that critics of EMH rarely make any effort to explain what it is for something (such as the CNS) to possess original intentionality. They appear to simply take it for granted *that* the CNS possesses original intentionality, without in any way offering an explanation for *why* this might be the case.

[22] For instance: (Sellars, 1956), (Bennett, 1964), and (Sterelny, 2003).

[23] Haugeland wrote, 'The trouble with artificial intelligence is that computers don't give a damn' (1979: 619).

[24] Our account of these three levels of self-criticism is drawn from (Haugeland, 2002a).

[25] This first level of self-criticism can get quite complicated. As Duhem and Quine have emphasized, there are cases in which it is not obvious which of two incompatible beliefs must be given up; either or both could be given up, given the possibility of broader revisions in one's beliefs about the world.

[26] As with the first level of self-criticism, identifying *which* norms work can get quite complicated. If one finds oneself regularly forming false beliefs it is sometimes not obvious which norms are to blame.

[27] The most detailed account of what it would be for a system to exhibit this third level of self-criticism is given in (Haugeland, 1998).

[28] As before, it is sometimes not obvious which norms those will be.

[29] It is worth noting that this makes Otto's notebook manifestly unlike most normal books or notebooks. Whether a particular string of text exhibits original or derivative intentionality is, we think, a question of how it behaves.

[30] At this point, more radical advocates of EMH might be inclined to say that we have not identified a relevant difference between Otto+notebook and Nat+cord, because on an expanded understanding of cognition, cognition is equally well spread throughout both such cases. After all, the *length* of the bungee cord should depend on *how high* the platform is from the ground and this length should be *revised* if it is too long or too short. We do not mean to dispute such a view; we only mean to dispute the view held by critics of EMH, according to which neither Otto+notebook nor Nat+bungee cord exhibit cognitive spread. If it helps to see what we take to be especially distinctive about Otto's case, contrast it with the (slightly ridiculous) imaginary case of Ben, who trips, falls off a cliff, and unintentionally wraps his leg around a loose vine, thereby creating a tightly coupled bungee-cord-like system, albeit one that is in no way the result of intentional human activity. Such a system does not exhibit cognitive spread of any sort. We take it to be importantly different from the Otto+notebook case in this regard.

[31] We think that Haugeland's example of the road to San Jose is similar to Otto's notebook in this regard. The key is to recognize that minds need not belong exclusively to individual human bodies. They can also belong to social entities, such as cities. It is some such larger social entity that takes responsibility for the proper functioning of the road in representing the way to San Jose. See (Haugeland, 1982).

[32] That said, we do think it is an important, outstanding problem (as noted in fn. 18 above) how to characterize the *different* contributions that different parts of cognitive systems make for the proper functioning of cognitive systems as wholes.

[33] For comments on previous drafts, we are grateful to Jake Browning, Nat Hansen, Eliot Michaelson, and Tyke Nunez.

References

Adams, F., & Aizawa, K. (2001), 'The Bounds of Cognition', *Philosophical Psychology*, 14.1: 43-64.

--- (2008), *The Bounds of Cognition.* New York: Blackwell.

--- (2010a), Defending the Bounds of Cognition. In R. Menary (Ed.), *The Extended Mind* (pp. 67-80). Cambridge, MA: MIT Press.

--- (2010b), 'The Coupling-Constitution Fallacy Revisited', *Cognitive Systems Research*, 11.4: 332-342.

Bennett, J. (1964), *Rationality.* London: Routledge & Kegan Paul.

Brentano, F. (1874/1995), *Psychology from an Empirical Standpoint*, (L. L. McAlister, Ed.). London: Routledge.

Clark, A. (1997), *Being There: Putting Brain, Body and World Together Again.* Cambridge, MA: MIT Press.

--- (2008), 'Pressing the Flesh: A Tension in the Study of the Embodied, Embedded Mind?' *Philosophy and Phenomenological Research*, 76.1: 37-59.

--- (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.* New York: Oxford University Press.

Clark, A., & Chalmers, D. (1998), 'The Extended Mind', *Analsysis*, 58: 10-23.

Crane, T. (1998), 'Intentionality as the mark of the mental', in T. Crane (Ed.), *Contemporary Issues in the Philosophy of Mind.* Cambridge: Cambridge University Press.

Fisher, J. (2009), 'Critical Notice of The Bounds of Cognition', *Journal of Mind and Brain*, 29.4: 345-57.

Fodor, J. (2009, Feb 12). 'Where is my mind? Review of Andy Clark's Supersizing the Mind', *London Review of Books*: 13-15.

Grush, R. (2003), 'In Defense of Some 'Cartesian' Assumptions Concerning the Brain and Its Operation', *Biology and Philosophy*, 18.1: 53-93.

Haugeland, J. (1979), 'Understanding Natural Language', *The Journal of Philosophy*, 76.11: 619-632.

--- (1982), 'Heidegger on Being a Person', *Nous*, 16.1: 15-26.

--- (1985), *Artificial Intelligence: The Very Idea.* Cambridge, MA: MIT Press.

--- (1990). 'The Intentionality All-Stars', in J. Tomberlin (Ed.), *Philosophical Perspectives, IV: Philosophy of Mind and Action Theory* (383-427). Atascadero, CA: Ridgeview Publishing Company.

--- (1995), 'Mind Embodied and Embedded', in L. Haaparanta, & S. Heinämaa, (eds.), *Acta Philosophica Fennica* (233–267).

--- (1998), 'Truth and Rule-following', in J. Haugeland, *Having Thought* (305-361), Cambridge, MA: Harvard University Press.

--- (2002), 'Authentic Intentionality', in M. Scheutz (ed.), *Computationalism: New Directions*. Cambridge, MA: MIT Press.

--- (2002), 'Syntax, Semantics, Physics', in J. Preston, & M. Bishop (eds.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (pp. 379-392), New York: Oxford University Press.

Hurley, S. (1998), *Consciousness in Action.* Cambridge, MA: Harvard University Press.

Menary, R. (ed.), (2010), *The Extended Mind.* Cambridge, MA: MIT Press.

Noe, A. (2004), *Action in Perception.* Cambridge, MA: MIT Press.

Rowlands, M. (1999), *The Body in Mind: Understanding Cognitive Processes.* New York: Cambridge University Press.

Rupert, R. (2004), 'Challenges to the Hypothesis of Extended Cognition', *Journal of Philosophy*, 101: 389-429.

--- (2010), *Cognitive Systems and the Extended Mind.* New York: Oxford University Press.

Searle, J. (1980), ‘Minds, Brains and Programs’, *Behavioral and Brain Sciences*, 3.3: 417-457.

--- (1983), *Intentionality.* New York: Cambridge University Press.

Sellars, W. (1956), ‘Empiricism and the Philosophy of Mind’, in H. Feigl, & M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science, Volume 1.* (pp. 253-329), Minneapolis: University of Minnesota Press.

Shapiro, L. (2011), *Embodied Cognition.* New York: Routledge.

Sterelny, K. (2003), *Thought in a Hostile World.* New York: Blackwell.

Thompson, E. (2008), *Mind in Life.* Cambridge, MA: Belknap.

Van Gelder, T. (1995), ‘What might cognition be, if not computation?’ *Journal of Philosophy*, 92: 345–381.

Varela, F. J., Thompson, E., & Rosch, E. (1991), *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Walter, S., & Kyselo, M. (2009), ‘Critical Notice of The Bounds of Cognition’, *Erkenntnis*, 71.2: 277-281.

Wilson, R. (2004), *Boundaries of the Mind: The Individual in the Fragile Sciences: Cognition.* New York: Cambridge University Press.

Wilson, R. A. (2010), ‘Extended Vision’, in N. Gangopadhyay, M. Madary, & F. Spicer (eds.), *Perception, Action and Consciousness.* New York: Oxford University Press.