Rehabilitating Representation

Brian Cantwell Smith[*]
University of Toronto[†]

1 Introduction

No concept has played a more important role in cognitive science than that of representation. The classic model of mind on which the field was founded was representationalist to the core, due to the second of its two founding assumptions. First, as reflected in its name,[2] cognitive science took intelligence to be epitomised by individual, rational, deliberative thought (in roughly Cartesian spirit, even though dualism was soundly rejected). Second, cognition so conceived was taken to consist in the formal manipulation of explicit, composite, language-like *representational* structures—rather as in logic. Although that founding view is often called 'computational,' for a variety of reasons I believe that that name is misleading,[3] so I will instead refer to it as **logicist**.

After being celebrated for many years, representation has recently suffered quite a drubbing. Challenges have been mounted

[*]Faculty of Information, University of Toronto
140 St. George St., Toronto, ON M5S 2G6 Canada

Last edited: October 20, 2009
Draft only (version 0.83) Comments welcome
Please do not copy or cite brian.cantwell.smith@utoronto.ca
Lightly-edited version of a paper delivered at a workshop on *Intentionality and the Natural Mind* sponsored by the Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis, March 19–20, 1998. A more thorough editing would generalize from cognitive science, in general bring the analysis up to date, and deal more substantially with notion of registration.
[2]I.e., as opposed to having been called 'the study of intelligence,' or some other moniker giving cognition less centrality.
[3]Smith, Brian Cantwell, 'One Hundred Billion Lines of C++', «ref».

on all sides—philosophical, neurophysiological, anthropological, and dynamicist. In its place, a spate of new views have been proposed, ranging from low-level neuronal models of brain function through autonomously navigating vehicular robots to high-level vaguely Heideggerian accounts of practice and sociality.[4] Though different in style and substance, these counterproposals are alike in one critical respect: they all recommend that the classic model be rejected in favour of a variety of more dynamic, embodied alternatives. Because of this common rejection of the classical model, these otherwise rather disparate alternatives are often loosely grouped together—originally under the label 'situated cognition,'[5] but more recently, perhaps because it more clearly incorporates neuroscience along with the other suggestions, under the label I will use here, of **embodied cognition**.[6]

The embodied alternatives tend to subject both founding assumptions to critique. First, instead of accepting as the premise that intelligence paradigmatically consists of cognition conceived as individual ratiocination, these views tend to privilege improvisational response and real-world (and, to varying extents, social) interaction—rather on the model of navigation. Second, there have been tendencies for all camps, each in its own way, to argue that the implementing mechanisms of this improvisational behaviour must be *nonrepresentational*.

As regards its status as socio-intellectual history—i.e., in terms of dominant rhetoric, prevailing assumption, and overall disciplinary profile—the shift from detached abstract reasoning to engaged material participation has largely been won. No one any longer denies the importance of context-dependence, of real-world interaction, of concrete embodiment. In fact contemporary students are likely to view favoring a logicist or 'formal symbol manipulation' view of mind to be as retrograde as holding a positive attitude towards pure introspectionism or Skinnerian behaviourism.

---

[4]«Refs»

[5]The session at the workshop during which this paper was first presented was entitled "*Intentionality and Situated Cognition*."

[6]Smith, Brian Cantwell, 'Situatedness/Embeddedness', Wilson & Keil (ed), *MIT Encyclopedia of the Cognitive Sciences (MITECS)*, Cambridge: MIT Press, 2001, pp. ■■–■■.

Independent of the merits of adopting a situated or embodied approach, however, it is not clear whether the wholesale embrace of **antirepresentationalism** may not ultimately prove as much of a straight-jacket as the original overly-zealous embrace of (especially 'formal') representationalism. If uncritically embraced as stand-alone directives, after all, even the most salutary correctives may lead down paths that miss their target as much as the views they were originally introduced to modify. Thus suppose Δ in figure 1 is introduced with the intention of shifting the target of mainstream inquiry (M) away from α and closer to β. If Δ is passionately embraced as a research path in its own right, instead of being recognized as an adjustment to M, it is likely to lead to α'—as far or farther from the desired β as the original α. And then if another correction Δ' to Δ is introduced in turn, the whole process may repeat, causing inquiry to proceed in a haphazard way. (Fundamentalism on the left is as untenable as fundamentalism on the right.)

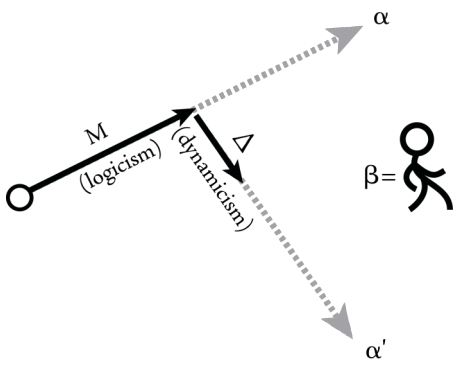

Figure 1 — Ideology in Cognitive Science

Something more straightforward is needed.

Perhaps the weakest suggestion is for a hybrid or **amalgamated** view: use non-representation wherever and whenever it works (empirically, pragmatically, theoretically), and then add in representation wherever it is appropriate or needed to handle "more complex" cases. Something of this sort is suggested in Clark's *Being There*,[7] and is advocated by as staunch an anti-classicist as Rod Brooks.[8] But no matter how commendably balanced, on its own that strategy is a bit vapid. Sure enough, as Braitenberg, neo-Gibsonians, and others have emphasized, non-representational

[7]Clark, Andy, *Being There*, Cambridge: MIT Press, 1998.
[8]Rodney A Brooks, 'Intelligence Without Representation,' John Haugeland, ed., *Mind Design II*, Cambridge: MIT Press, 1997, pp. 395–420.

mechanisms are capable of producing vastly more complex and subtle behaviours than classicists ever imagined.[9] But a simple amalgamation strategy doesn't answer any of the constitutive questions: when or why representation might be needed, what contributions it may (uniquely?) be capable of supplying, when it is *not* required or advisable, etc.—to say nothing of what the powers and limits might be of pure mechanism or pure embodied behaviour.

Moreover, to assume that the two traditions can be glued together without alteration—as if in an assembly—is a bit of a dream. The suggestion also fails to illuminate the question of what kind of representation would best suit a combinatorial approach (abstract, formal, logical, imagistic, etc.); nor does it say anything about what should play the role of anti- or nonrepresentational complement (physical dynamics, existential thrownness, etc.).

We need to cut deeper.

A more powerful idea is suggested in figure 2: what I will call a **generalisation** strategy. Rather than assume that logic encapsulates the essence of what it is to be representational, the suggestion is to recognise representation as an (at least potentially) richer and more encompassing notion in its own right, and then to identify (and perhaps criticize) the logicist variety as just one particular species. Among other merits, this approach has the virtue of not "giving away" the notion of representation to the predecessor view, as if logicians somehow understood representation's be-all and end-all.

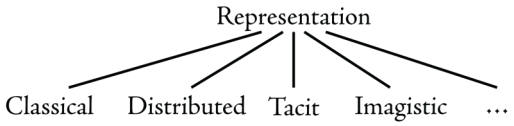

Figure 2 — Species of Representation

From an intellectual point of view, the generalisation strategy requires dissecting the traditional conception of representation into two parts: (i) what is universal about representation in gen-

[9]The point was made as early as in Herb Simon's *The Architecture of Complexity*. Cambridge, MA: MIT Press, 1969. See also Braitenberg, Valentino, *Vehicles: Experiments in Synthetic Psychology*, Cambridge: MIT Press, 1986; other «refs».

eral, applicable to all species; and (ii) what is specific to the particular form of representation embodied in the classical view. In practice the strategy has rarely been approached so theoretically. It has instead proceeded in a more "bottom-up" way, through numerous attempts to identify other (allegedly non-logical) species of representation. Of many suggestions, perhaps three are most famous: (i) *imagistic*, *iconic*, *pictorial* or *visual* representation—a perspective from which logicist representation is viewed as fundamentally *linguistic* or *propositional*;[10] (ii) *procedural* representation—in contrast to the presumptively *declarative* character of representation in logic;[11] and especially since the rise in popularity of connectionist and other network models, (iii) *distributed* representation—as opposed to what was most often in these debates simply called *classical*.[12]

Since they were largely operating within a representational context, the defenders of these alternatives tended to concentrate on bringing particularities of specific cases into focus (logicist and other), rather than addressing the overarching issue of representation in general. The result was to leave us without much of an understanding of how representation might or should be combine (in an intelligent agent) with more direct forms of dynamics or embodiment. But there were a number of significant exceptions, perhaps especially including John Haugeland, who not only attempted to compare and contrast what he calls *logical*, *iconic*, and *distributed* "genera", but who also made some remarks about the general case.[13] Strikingly, he introduced his paper as "even more than usually tentative and exploratory"; called its results "at best preliminary and incomplete, perhaps much worse";[14] and took up his discussion of representation-in-general with yet an additional caveat:[15]

[10]«Ref Kosslyn, Shepherd, and others»
[11]'«Ref Winograd and others»
[12]«Ref the PDP volumes, the Smolensky/Fodor debates, etc.»
[13]Haugeland, John, "Representational Genera," in W. Ramsey, S. Stich & D. Rumelhart (eds.), *Philosophy and Connectionist Theory*, Hillsdale: Lawrence Erlbaum, 1991, pp. 61–89. Reprinted as ch. 8 in Haugeland, John, *Having Thought*, Cambridge, MA: Harvard Univ. Press, 1998, pp. 171–206.
[14]op. cit.; p. 172.
[15]op. cit.; loc. cit.; emphasis added.

> "An explicit account of *representation as such* will not be necessary; that is, we can get along without a prior definition of the 'family' within which the genera are to be distinguished. A few sketchy and dogmatic remarks, however, may provide some useful orientation, as well as places to hang some terminological stipulations."

Yet in spite of his cautionary remarks, the three paragraphs that Haugeland devotes to the topic not only contain substantial insight, but are also so widely cited as cognitive science's best char-

Haugeland on Representation[†]

"A sophisticated system (organism) designed (evolved) to maximize some end (such as survival) must in general adjust its behavior to specific features, structures, or configurations of its environment in ways that could not have been fully prearranged in its design. If the relevant features are reliably present and manifest to the system (via some signal) whenever the adjustments must be made, then they need not be represented. Thus, plants that track the sun with their leaves needn't represent it or its position, because the tracking can be guided directly by the sun itself. But if the relevant features are not always present (manifest), then they can, at least in some cases, be represented; that is, something else can stand in for them, with the power to guide behavior in their stead. That which stands in for something else in this way is a *representation*; that which it stands in for is its *content*;[a] and its standing in for that content is *representing* it.

"As so far described, 'standing in for' could be quite inflexible and ad hoc; for instance, triggered gastric juices might keep a primitive predator on the prowl, even when it momentarily loses a scent—thus standing in for the scent. Here, however, we will reserve the term 'representation' for those stand-ins that function in virtue of a general *representational scheme* such that: (i) a variety of possible contents can be represented by a corresponding variety of possible representations; (ii) what any given representation (item, pattern, state, event, …) represents is determined in some consistent or systematic way by the scheme;[b] and (iii) there are proper (and improper) ways of producing, maintaining, modifying, and/or using the various representations under various environmental and other conditions. (This characterization is intended to be neutral not only among genera, but

[†]Haugeland, John, "Representational Genera," pp, 172–73. Emphases and notes in the original.

acterisation of representation that I have taken the liberty of reproducing them here, to serve as something of a starting point (sidebar, p. ■■).

In order to give the generalization strategy as much play as possible, the characteristics of logical representation identified by advocates of alternative species—i.e., its being *linguistic* or *propositional*, *declarative*, *systematic* and *productive*, etc.—can be added to the logicist characteristics most often criticized by the anti-representationalists: the fact that logic is allegedly *explicit*, *formal*, *context-independent*, *static*, and *abstract*; the fact that it emphasizes

also between internal and external representations, and between natural and artificial schemes.)

"Since the content of a given representation is determined by its scheme (and since the point of the facility is to be able to represent what isn't present or currently accessible), it is possible for representations to misrepresent. What this amounts to will vary with the specific scheme, and even more with its genus; but it must hark back eventually to the possibility of the system(s) using it being *misguided* in their attempted adjustments to the features of the world. But misrepresentation should not be confused with improper deployment on the part of the using system, nor bad luck in the results. These can diverge in virtue of the fundamental holism underlying what can count as a representation at all: the scheme must be such that, properly produced and used, its representations will, under normal conditions, guide the system successfully, on the whole. In case conditions are, in one way or another, not normal, however, then a representing system can misrepresent without in any way malfunctioning."

a. This use of the term 'content' is not altogether standard. Most contemporary authors (and I, in the other essays in this volume) mean by the "content" of a representation something distinct from the object it represents, and which determines that object (as sense determines referent, for instance). Here, however, I mean by 'content' that which the representation represents—the "object" itself—but as it is represented to be (whether it is that way or not). Thus, it is a possible object—which may in fact be actual, or similar to something actual, or neither. [Note added 1997.]

b. For instance, if (or to the extent that) particular representations are tokens of well-defined types, the scheme will determine the content of any given token as a function of its type—or, at least, these will determine how that content is determined. Thus, if any extra-schematic factors (such as situation or context) co-determine contents, then which factors these are and how they work are themselves determined by the scheme and type.

reference, *rationality*, and *truth* over other semantic properties and norms deemed by some to be more appropriate to pragmatic intelligent conduct; etc. Unless any of these properties can be defended as constitutive of representation itself, the generalist would want (i) to forge a notion of representation that is not committed to them, and then (ii), if they are not only inessential to representation as a whole but also inappropriate for the full range of cognitive behaviours, to identify other species that, while still genuinely representational, do not exhibit those specific characteristics of logicism.

Though not explicitly described in these terms, support for such a generalising approach can be found in a flurry of recent discussions of representation in the philosophy of mind.[16] The strategy has also had the benefit of leading cognitive scientists to read in areas of philosophy beyond logic and the (relatively narrow) classical "Language of Thought" school of philosophy of mind—e.g., to look to Ryle, Merleau-Ponty, James, Heidegger, Dewey, Langer, etc., for inspiration.[17]

Note too that generalisation can easily be added to amalgamation in a combined hybrid strategy. There is no need to insist that a representation, even appropriately generalised, must apply to *all* aspects of human cognition. The point is just to make room for the possibility that some (or perhaps even many) aspects of intelligent behaviour may require some notion of representation for their proper explanation—i.e., to recognise that there may be aspects of cognitive behaviour that cannot be accounted for by (for example) a purely dynamical approach, even if they do not fit into the classical "logicist" framework.

Read this way, the generalisation strategy has much to recommend it, and in many ways I will adopt it here. But it, too, especially by itself, does not cut deep enough.

In this paper I will argue for a third approach—something I will call a **reconstructionist** strategy. The (admittedly *ex post facto*) argument for reconstruction runs roughly as follows:

[16]«Ref Cummins, Chemero, Clark and Grush, etc.»
[17]«Refs»

1. It is true that the classical model is too specific (narrower) than is required or appropriate for many of cognitive science's purposes.
2. It may also be true that some (even constitutive) aspects of a person's overall cognitive processes may be nonrepresentational—as suggested in the amalgamationist strategy.
3. Independent of the merits of (2), however, it is also true that the logicist model of representation is narrower than representation per se requires, and so the logicist approach to representation should be generalised, and new non-logicist species of representation identified and explored—as recommended in the generalisation strategy.
4. But something stands in the way of our doing this generalisation.
5. Although the classical model was based on some very deep insights into the nature of representation,
6. Those insights were expressed in ways that were not only too narrow, but in addition outright misleading—i.e., not just false of representation in general, because restricted to the circumstances (and expressed in the language) of the specific view, but inadequately understood *even in that restricted (classical) case*.
7. What we need, therefore, is not just to generalise, but to reconstruct, the classical view: reframe and rephrase it, re-understand its essential features.

For a simpler but different example of reconstruction, to see the strategy in action, consider a case I will talk more about below: the constraints of "computational effectiveness" that lie at the very basis of logic and computer science. There is no more important conceptual ingredient in the classical view than this notion of what can be algorithmically or mechanically done.[18] For various reasons, as we will see, these effectiveness constraints, even in syntactic guise, have classically been viewed as *mathematical* and *abstract*. What I will argue is that even in classical settings, and in

[18]The term 'effective' is inscribed in the foundations of computer science: its core theory is called a theory of *effective computability*.

spite of the character of classical analysis, they are not, in point of fact, abstract after all, but instead are direct (if implicit) consequences of the material character of the underlying computational substrate. They have been *understood* as abstract, but classical understanding is wrong. In point of fact they are *concrete*.

This is an example of reconstruction, not generalization, because I am not claiming: (i) that effectiveness can be legitimately understood as abstract in the classical case—i.e., in situations where formal logical explicit representation or inference is mandated; but (ii) must be understood as concrete (material, physical) in more general situations—e.g., those involving non-classical forms of computational and/or representational activity. Rather, I am making the stronger claim that even *in paradigmatic cases of first-order logical inference*, the operative constraints on "what can be done" (what can be proved, what can be inferred, what can be mechanized, what can be computed) are and always have been ultimately physical, even if they have not classically been understood in that way. In others words, the reigning theoretical presumption that effectiveness and computability are appropriately understood abstractly or syntactically isn't too *narrow*. It is *false*.

In what follows I will to varying degrees adopt all three strategies—amalgamation, generalization, reconstruction—but in reverse order:

1. First we need to reconstruct the classical view, which among other things will allow us to see, in some depth, what was particular about the classical view, and what circumstances if any recommend its use;

2. Then we will be able to generalise the notion of representation appropriately, formulating a more powerful, encompassing replacement;

3. Then—and only then, with a generalised notion in hand—we can address the question underlying the first amalgamationist strategy: of which aspects of cognition do, and which aspects do not, need to be understood in representational terms;

4. Once that is in hand as well, we will have arrived at a possible *substrate* for a comprehensive account of mind.[19]

Three final preparatory comments.

First, it is ironic that representation has been misunderstood on *both* sides of cognitive science's pro- and antirepresentationalist debate, blocking substantive progress. But although it helps to point this out, my aims are not ultimately critical. Rather, what I want to figure out is how to be *positive about both sides at once*—i.e., how to do justice to the intuitions underlying each. The issue is not merely rhetorical or motivational—or even socio-intellectual, where (as mentioned) the issues are largely settled. Like most modern writers, I approach cognition sympathetic to a renewed emphasis on embodiment, activity, and practical "being in the world," of the sort that motivates the embodied cognition movement. At the same time, however, I am concerned that many of the profound insights that underwrite the classical model (particularly, as we will see, semantical insights) are being lost, in the rush to embrace "in the world" concrete embodiment.

More pointedly—and in a sense this is the real aim of the paper—I worry that, in eschewing abstract formality in favour of concrete materiality, a spate of embodied cognition theories, from cognitive neuroscience to cultural theory, even if dressed in impeccable scientific credentials or urbane French garb, are unwittingly falling prey to a kind of *causal reductionism* or *causal fundamentalism* incapable of understanding what is ultimately distinctive about minds and mentality—having critically to do with semantic directedness. Put it this way: the most urgent challenge for embodied cognition, in my view, is to

> *Preserve—perhaps even rescue—semantics through a (beneficial) shift from abstractness to concreteness.*

It will take some work to see what this comes to. I will start with two critical reconstructions, followed by a dozen or so targeted

[19]It is no theory of mind; that would be something vastly more ambitious. I call it a possible substrate only in the hope that, by diagnosing the relation between physicality and concrete representation, it may supply conceptual terms in terms in terms of which a successful theory of mind might be formulated.

generalisations. Once we have those in place, we will be able to start combining what matters about each side of this overly dichotomised debate into a unified and durable successor.[20]

Second, it is important to understand that the pro and antirepresentational debate in cognitive science falls on the **sub-personal** side of the *personal/sub-personal* distinction.[21] Assume that by 'subpersonal' I will refer to the mechanisms or ingredients out of which intelligent creatures are made, and by 'personal' will refer to the full-blooded intentional agents thereby physically constituted. It is the full person, that is, who is the subject of consciousness, the bearer of rights, the participant in social norms, the member of community. It is the subpersonal mechanisms that implement or realize persons with which cognitive science is primarily concerned. I would thus take as falling within the scope of the questions being addressed here debates about the representational character of the retinotopic map in areas V1 through V5 of the visual cortex, and debates about whether, in empathy, we *represent* the emotional lives of others, or do something more akin to *taking them on*. But I would not, by itself—at least not without comment and consideration—take a positive answer to either question to be evidence that, *as people* (i.e., at the personal level) we "represent" our environment or our friends in the course of our daily lives.

By making a personal/sub-personal distinction—by foreswearing an *identification* of people with the meronomic components of their bodies—I absolutely do not want to suggest that the two are *independent*. Neither do I want to endorse claims, such as those of McDowell, that attribution to ingredient mechanisms of such intentional characteristics as "being representational" is merely "as if".[22] On the contrary, I believe that the relation between the representational, semantic, intentional and/or normative character of

[20]Interestingly, this recombinant reconstruction is necessary in order to achieve another goal of great importance: that we unify the understanding of representation that serves in technical fields (such as logic, computer science, linguistics, cognitive science, etc.) with understandings of representation in literature, the arts, and humanities.

[21]«Refs; including McDowell's response to Dennett»

[22]«Ref»

"that of which we are made" and the representational, semantic, intentional, and normative character of "we who are thereby made" is extraordinarily vexed. In other contexts I will argue that co-constituting ties bind the authenticity of the full-blooded intentional involvement of persons in the world and the genuineness of the representational character of the material ingredients of the world in virtue of which are they are such full-blooded participants.

But this is not the place to pursue such questions. Here I want simply to introduce a term that I have developed more fully elsewhere, which will help us to mark the personal/sub-personal distinction and to stay out of the notorious conceptual confusion that stems from ignoring it. In particular, *unless explicit comments are made suggesting otherwise*, I will say that, at the personal level, we **register** the world in terms of the objects, properties, situations, states of affairs, features, etc., that we thereby take it (the world, that is) to consist in. Thus as I write these sentences, as it happens, I *register a building across the street*, *register a lake on the horizon*, *register a thorny academic situation of which I have just learned as petty and unfortunate*, etc.

A few comments about the notion:

1. By 'registration' I intend to index the fact, shared by representation, that human thought, perception and understanding of the world is ineliminably 'as'.
2. I take the term to be neutral as to any distinction between or among *sense*, *perception*, *thought*, *judgment*, etc.
3. Unlike 'conceive' or 'cognize' (and in this respect more like 'see' and 'perceive'), I take 'register' to be a "success" verb. If, in ordinary circumstances, a person registers—i.e., takes there to be—a tree, then it is fair to assume that there was a tree there to be so taken, and that the person did so take it, in the full semantically and normatively appropriate way.[23]

[23]This is a rather realist characterisation of "success"; it should be replaced as appropriate for other metaphysical views. The point is simply that "α registered β" should be true just in case something roughly of the form "There is β and α took it to be β" is true.

4. The term 'register' is usefully neutral on the division of responsibility between person and world for the resulting ontological "take"—i.e., is neutral as between naïve realism (I successfully register a table *as* a table because it *is* a table), strong forms of constructivism (I successfully register it as a table because of the contingent and historical forces constituting the social community of which I am a member, or even due to the particular exigencies of my own individual history), idealism, solipsism, and many other epistemic, ontological, and metaphysical proposals.
5. As should be evident from the above (including occasional use of the qualifier 'successful'), I take registration to be normatively laden, in the philosophical sense of serving as the subject of such issues as truth, objectivity, worth, etc.
6. By following the verb's direct object with 'as …', the construction facilitates at least a first step towards distinguishing how we, as theoreticians (cognitive scientists, epistemologists, etc.), register situations or phenomena in the world, and how we take them to be registered by other subjects or people or agents, of whom we may be speaking. Thus I might say, of an infant, "She registers her mother's coming to the door not as the re-appearance of a recognized individual object, but more as "re" or "repeating" placement (in Strawson's sense) of the feature *Mama*."[24]
7. If *not* used with an explicit 'as' construction—i.e., notwithstanding (6)—I will assume that the direct object of 'register' to include the aspectual nature of the way in which that phenomenon or entity is registered by the (individual designated by) the sentence's subject. In this way 'register' differs from at least common uses of such perceptual verbs as 'see.' Thus while some would claim that it is possible for the sentence "Randy saw the Northern Lights" to be true even if Randy did not recognize them *as* the Northern

[24]It is only a first step, because of the evident but fraught issue of *how we register the subject's registration* (e.g., in the 'γ' part of the sentence 'α registers β as γ')—including whether we can, and if so how much, and in what respects.

«For the infant case, ref Jun's Duke dissertation.»

> Lights, I consider it an implication of the sentence "Randy registered the Northern Lights" (without any following 'as' clause) that Randy did so take them.[25]

In these terms, I would characterize the **representational theory of mind** as a theory that claims that *human cognition is underwritten by processes involving the manipulation or use of representational ingredients*. *Per se*, that is, the computational theory of mind does not mention registration; if it needs to explain registration (as I believe it must), then it must do so as a consequence of the substantive claim that registration is something that people do. I thus consider it to be a substantive question how much of human existence and/or participation in the world rests on registering it—as opposed, say, simply to bumping into it, or responding as a purely physical or mechanical device. Similarly for two questions to which in the long run the present investigation is likely to be primarily relevant: (i) how much human registrational capacity is underwritten by representations—either internal, external (as suggested in discussions of scaffolding etc.), communicative, etc.;[26] and (ii) conversely, whether representations are implicated as ingredient mechanisms to underwrite human capacities that, at the personal level, do not involve registration.

For now, it is enough to say that when speaking at the personal level, of whole human beings, I will speak of *registration*, unless (in which case it will be explicitly marked) it is genuinely personal-level representation that is at issue, as for example might arise in a discussion of parliamentary democracy. Except in such marked cases, however, uses of 'representation' will refer to entities that are constitutive, realizing, ancillary, external, supportive, communicative, or otherwise implicated in the world that we as persons inhabit.

Third and finally, although the paper is entitled "Rehabilitating Representation," it is the *notion* of representation I aim to renew and refurbish, not the representational theory of mind.[27] Indeed,

[25]I.e., the direct object position of the verb 'register' is thus not assumed to be referentially transparent.
[26]«Ref Clark and others»
[27]If renovating concepts ruffles your ontological or epistemological feath-

it is no part of my purpose here to argue for or against such a representational theory. My concern is only that representation is a more powerful notion than recent treatments would lead one to suspect. The question of whether the mind is representational strikes me as both substantial and open—a question to which we are as yet far from knowing the answer.

ers, take this as elliptical for refurbishing discourse that makes substantial use of the concept.

2 Logic

The concretization of effectiveness described in §1 is just the first step in our reconstruction of the classic logicist view. Others steps have to do with semantics, formality, and the structure of norms. To understand any of them, we need a clear grasp of the conceptual (though not technical) structure of logic—the aim of this section. I will assume a modest working familiarity with basic logical notions, of the sort presumed throughout cognitive science; my aim here is simply to clarify some of logic's underlying conceptual framing.

In particular, I will proceed in two steps: (i) describing an atemporal or static logical or representational basis, and (ii) a computational increment, introducing the notion of process.

2a Logical basis

As diagrammed in figure 3, the classic logical picture consists of five ingredients, grouped into three kinds:

1. Two realms: one syntactic (S) and one semantic (D);
2. Two relations, one on each realm: a "proof-theoretic" derivability relation P on S,[28] and a real-world or domain-theoretic entailment or dependence relation R on D;[29] and
3. A semantic interpretation function (I) from S to D.

The syntactic realm S consists of the representations themselves—

[28]Often written as an infix '$\vdash$', as in '$S_1, S_2 \ldots S_i \vdash S_k$'

[29]Entailment ('$\models$') is usually understood as a relation on S, as in '$S_1 \ldots S_i \models S_k$', or as a relation among elements (or sets of elements) of D and a sentence S_i, as in '$D_i \models S_k$' (in for example a case where D_i was a possible world in which S_k is true). What I mean by saying that entailment (R) is defined on D is that, however it is formally defined, entailment ultimately rests on a relation R defined among elements of D, to which it relates, in any sense in which sentences are involved (except self-reference) through I. It is R, the relation among elements of D, that is, that "wears the trousers" as regards entailment. For example, suppose one says that that S_1 (in S) entails S_2 (in S)—i.e., that $S_1 \models S_2$. That would be true just in case the interpretation $I(S_1)$ *bears* R to the interpretation $I(S_2)$.

In a standard extensional model of first-order logic, R would be something like inclusion, where I maps sentences onto sets of models in which S is *true* (i.e., so that $S_1 \models S_2$ just in case $I(S_1) \subset I(S_2)$).

structures it is in general convenient to call **representational vehicles**: typically, expressions in a formal language. Paradigmatically, elements of *S* are formed from a finitely-specified set of atomic elements or *simplexes* (variables, constants, predicate and relation symbols, quantifiers, various sentential operators, and so on), assembled in a variety of ways into well-formed *complexes* (sentences, quantified terms, etc.) via inductively-specified syntactic composition rules, in the familiar way. Classical knowledge representation schemes in artificial intelligence (which I am including in the general "logicist" camp) introduced a bevy of structural and aesthetic variations. Various properties specific to *written* languages, for example, having to do with one-dimensional lexical syntax (including for example the idea of *named* variables) were set aside in favour of a more abstract conception of representational structure, leading to proposals that more closely resembled graphs, or even abstract structures from non-well-founded set theory.

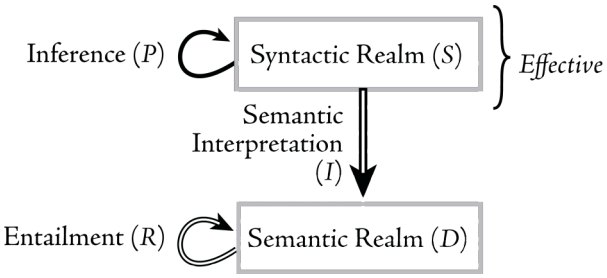

Figure 3 — The Conceptual Structure of Logic

Details do not matter here, though, since our aims are conceptual rather than technical. It is enough to note that in classical models, syntactic structures (representational vehicles), both simple and complex, are assumed both to be definable and identifiable on their own—i.e., to have determinate, autonomous identity conditions, independent of and explanatorily prior either to the semantic realm, or to the two relations (derivability or interpretation).

The semantic realm *D*, on the classical view, is normally treated model theoretically, and in that guise taken to be *abstract*. Again, elements of the semantic realm are also (usually) taken to be discrete and determinate, with ontologically and explanatorily autonomous identity conditions. In a typical case, the semantic realm *D* would be assumed to consist of a (possibly set-theoretic) domain or structure of objects, properties, relations, functions,

etc.—again, in a wholly familiar way.

The proof theoretic or inferential relation *P* is defined over the syntactic realm *S*. *Legitimate* inference relations *P* (out of the space of all possible *P*s) are identified in virtue of various *semantic* constraints on *P*, defined in terms of *I* and *R*, as we will see. It is they, ultimately, that give *P* its main substance. But before semantics is allowed to get a a toe-hold, *P* must satisfy three critical conceptual well-formedness conditions.

1. *P* must be definable over the *formal* or *syntactic* properties of the representational vehicles (i.e., the elements of *S*).

Stunningly, what it is to be a formal or syntactic property isn't entirely clear;[30] what it is to be a syntactic property is rarely theorized. Nevertheless, as made famous in cognitive science circles through Fodor's formulation of his **formality condition**, *form* or *syntax* is generally taken to have both a positive and a negative aspect (sidebar). Positively, it has to do with the grammatical or syntactic structure—namely, those properties, including the identity conditions, of the elements of *S* in terms of which *S* is defined; negatively—and this is critical—syntax is assumed *not to involve or make reference to any semantic properties*. Thus it would be malformed, because not formal, to define an inference relation that applied only to *those expressions that Jerry Fodor currently favours*, or to *those expressions that are true*.

2. In a computational or cognitive context, the derivability relation *P* must also be *effective*, in the sense of being able to carried out, or at least checked, "mechanically."

The rule "From expression s_1 derive the constants '*T*' or '*F*' depending, respectively, on whether, a hundred years from now, S_1 will or will not have appeared more often in published logic textbooks" is adequately formal by the first criterion, but fails to be effective by the second.

[30]Since logics are usually introduced individually, by ostension, the syntactic properties of a particular system are usually simply pointed out, and accepted, by-passing the requirement for a general account. But see below.

Fodor's Formality Condition†

"What makes syntactic operations a species of formal operations is that being syntactic is a way of *not* being semantic. Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference, and meaning. Since we don't know how to complete this list (since, that is, we don't know what semantic properties there are), I see no responsible way of saying what, in general, formality amounts to. The notion of formality will thus have to remain intuitive and metaphoric, at least for present purposes: formal operations apply in terms of the, as it were, shapes of the objects in their domains."

†Fodor, Jerry, "Methodological Solipsism," «Ref»

3. In rather anti-Wittgensteinian spirit, syntactic properties, as well as having to be effective and non-semantical, are required, in logical settings, to be both syntactically and semantically defined *without regards to their use*.

Thus no room is made for defining a relation Q that is transitive so long as it is not used more than three times in a derivation.[31]

Given well-formed syntax and appropriate compositional rules, the interpretation function I is typically defined inductively in the following compositional sense: given a complex (syntactic expression) S^* of S, consisting of parts S_1, S_2, S_3, etc. the interpretation $I(S^*)$ is assumed to be defined in terms of the interpretations $I(S_1)$, $I(S_2)$, $I(S_3)$, etc., by an inductively-specified process of formation that is purely a function of S^*'s *formal* (in the positive sense—i.e., *grammatical*) structure. The inductive structure of the syntactic formation rules, plus this so-called **semantic compositionality** (essentially: an isomorphism or homomorphism between the grammatical structure of S and the "formation" of in-

[31]I.e., so as to license the inference $Q(x,y)$ & $Q(y,z) \Rightarrow Q(x,z)$ up to three times per derivation, but no more—as one might be tempted to suggest for a relation such as *Near*. Of course this constraint (any many others) can be "worked around" by coding the number of applications in varieties of the predicate itself, but the rule stands that, per se, use is not a legitimate ground for syntactic definition.

terpretations D[32]) ensures that the language is *systematic* and *productive*—capable of expressing untold new things, in a regular way. It is normally of great importance that all the basic vehicular ingredients—the stock of elements comprising S, the grammar or syntactic formation rules by which they are assembled into complexes, and the interpretation function I (that is: everything except the semantic domain D) be *finitely specifiable*. "Infinite expressive power via finite means" is something of a mantra in logicist quarters. It is all a little a fantastic Meccano or Erector set, with an unlimited supply of perfect, infinitely strong, weightless parts, in a world without friction, rust, or decay.

2b Computational component

Needless to say, for even the most cursory account of logic a vast amount more needs to be said—about truth, for example, and soundness and completeness. I'll make a few such remarks in a moment. More important for our purposes, however, is how much has not, *and does not need to be*, said: (i) anything about the nature of the representational vehicles, for example—whether they are linguistic, pictorial, distributed, etc.; (ii) anything about the nature of the domain D—such as whether it is composed of fields, features, objects, properties, dreams, ideas, or such. Logicism's specific assumptions in this regard will come up later, in the generalization phase; one of the points of framing the conceptual structure as we have done is to prepare it for a much wider than normal set of possibilities

Instead, consider what is involved in turning the foregoing logical picture into an active, computational system—of the sort that cognitive science classically imagined to be an appropriate or at least possible model of intelligence.

If not actually static, representational systems of the sort just described are at least *a-temporal*; the proof or derivability relation (P) is just that: a formally specified abstract *relation*. But even in logical guise there is almost always a residual bias towards think-

[32]Note: nothing requires that D itself have any structure whatsoever. So for example, the interpretation of the (inductively-defined) expression ((2+3)*(4/5)) is the atomic number four; not anything with a structure corresponding in any way to the grammatical form (_,_)·(_,_)).

ing of *P* in a forward direction, as bearing some relation to active patterns of rational thought. In the hands of cognitive science, given its interest in how people actually work, background bias becomes foreground concern: it is necessary to convert *P* into a *temporal process*. This "temporal mechanisation" of *P* can be viewed as the "computational turn" on the logical framework.

Needless to say, mechanising inference is far from trivial. Some raw materials are already in place: the relation *P*, and the formal properties of the expressions of *S* in terms of which it is defined, are all already constrained to be formal/syntactic, and so *P* (it is understood) is thereby amenable to direct computational implementation. The major problem is that, for any plausible representational system of the indicated sort, the proof relation *P* will be wildly branching. Given a set of expressions S_1, S_2...S_k, it becomes a major issue to determine which particular S_i to have the system produce, out of the (typically vast) set licensed by *P*. Many knowledge representation, planning, theorem proving, "search strategies," and other cognitive science projects can be seen as attempts to solve this problem, under the general rubric of "controlling inference".

The standard way to attack this problem was the following: one implements the representational vehicles—the representational vehicles, the elements of *S*—as data structures in a computer system, and then writes a program (we'll call it PROG) whose function is to make transitions from initial elements of *S* to final elements of *S* in some interesting or plausible way. The idea, that is, is that program PROG specifies the (potentially quite complex) behavior of the inference process *over the specified domain of representations S*. In computational jargon, this can be characterised as saying that the program specifies the behavior of a process over the data structures.

A note in passing. Though there is nothing especially problematic about proceeding in this way, one fact about this situation has proved remarkably distracting. It has to do with a conflation of, and subsequent theoretical confusion between, two distinct languages. The standard way to construct representational vehicles, as mentioned above, is to construct them in terms of a representational system which I will call the **representation language**

(L_S—i.e., the language in which syntactic formulae *S* are defined). The programs for controlling inference, at least in their so-called "source" versions, typically consist of a set of expressions in *another* formal language, which we will call the **programming language** (L_{PROG}). (For example, suppose one were to implement an inference system to work with expressions in the first-order quantificational calculus: the programming language might be C++, whereas the representation language would be the an encoding of first-order quantificational calculus expressions in C++ data structures.

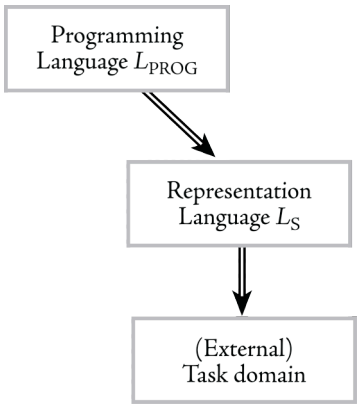

Figure 4 — Programming

As indicated in figure 4, what matters is that L_S and L_{PROG} are *different languages*. They refer to different domains, are subject to different constraints, and have radically different interpretations. One simple way to characterise their relation is to note that the programming language L_{PROG} stands at one level of semantic ascent *above* the representation language L_S: what expressions in L_{PROG} denote (i.e., are "about") is *formal transitions over (syntactic) elements of* L_S. The programming language L_{PROG} is a meta-language, that is; representation language is object language. It is only because the two languages have been confused, I believe, that cognitive science has called the classic representational view of mind *computational*.[33])

Sometimes, the process that makes the moves from starting expression to final expressions (perhaps via a long series of intermediate steps) is reified into a separate conceptual component of the overall system—leading people to say that the representations are "read" and "written".[34] However identifying the specified process as a sub-process of the overall behavioural system is both unnecessary and generally confusing[35]—and anyway there is no

[33]Smith, Brian Cantwell, 'One Hundred Billion Lines of C++', «ref».
[34]E.g., cf. Haugeland.
[35]In part because there is a tendency to confuse the process over representations (i.e., the process that manipulates the sub-personal representational structures *S*) with the overall process of which it is a part (which, if

reason why the overall system has to be implemented in that way. More generally, therefore, and as far as possible to avoid confusion, I will avoid any talk about programs PROG and the programming language L_{PROG} in which they are written, and simply talk about one process (the overall, inclusive, one, of which the representational vehicles are a part), whose behavior comes about as the result of **effective transitions** from elements of s to new elements of s.

2c Semantics

Given this overall picture of logic, four points need to be highlighted, to prepare us for subsequent reconstruction.

The first has to do with the question of just which parts of figure 3[36] are, and which parts are not, subject to effectiveness (computability) constraints. The answer is straightforward: effectiveness concerns only the upper third of figure 3; the formal (in the positive sense) properties of the representational vehicles S, and the proof-theoretic or inferential relations P between and among them—i.e., just 2 of the 5 ingredients of the overall logicist picture.

Crucially, not only is there no requirement that the interpretation function I be effective; there is no reason to believe that thinking that I must be (or even is) effective is even conceptually coherent.[37] For "effective" means something like "mechanically implementable," or "is a kind of operation that a Turing machine can do"—i.e., a temporally extensive operation that starts and ends with concrete entities or at least arrangements—a type of constraint that doesn't make any sense when talking about the non-temporal relation between, say, a numeral and a presumably abstract number. Moreover, what is *formal*, in what we are calling the (positive) grammatical sense, is again just the representational vehicles. The semantic domain D is sometimes characterised as formal, but whatever that means, it must be in a different sense from either of the positive or negative readings we have given to

---

the cognitive model or AI project were successful, would be a person). This is just one of the mistakes that Searle makes, for example, in his famous Chinese Room thought experiment «refs».

[36] P. ■■.

[37] Cite AOS, appropriate volume.

that term (for example: it might mean *mathematical*).

More generally, whatever is the nature of the interpretation relation *I* between representational vehicles and the entities they designate or denote, it is not something that *happens*. The numeral '2' designates the number two, or so at least it is normally presumed; but that "designation" is not a *process*, not something that *happens*, not something that takes *energy* or *time*. Semantics—at least in the small—is something that "obtains."

Thus to take a human example, suppose one has a thought about the Brooks Range, or about Cheops, or about the great day on which the United States elects its first female President. The relation between one's thought (be it a state of mind, an active brain process, or whatever) and what it is about—the "directed arrow of reference" that starts in one's head and leaps out across time and space to the north slopes of Alaska, to an Egyptian ruler in the third millennium B.C.E., to a day in the (with luck) not too distant future—that referential arrow is not part of the energetics of the world. Referential entities, even referential *activities*, do not bathe their referents in any flux of discriminable energy. As I have said in another context, not even the NSA[38] could build a meter, to be worn in one's pocket, that could detect whether the bearer was the subject of an intentional act. The problem is not that referential signals are too faint, or that our physics is not sufficiently advanced, or that some form of quantum mechanical wizardry is at work. Rather, the reason is that semantic properties—being referred to, being true, being consistent, etc.—*are not effective*. As we will see, that is one of their enormous virtues—something that causal reductionists ignore at their peril. (And of course it is a good thing that reference, for example, is not effective; it is exactly the fact that reference is *not* effective that allows us to refer to the past, or to refer to the future, without therein violating physical proscriptions on forward or backwards causality).

2d Naturalisation

So that's the first point: syntax and inference (proof) are subject to effectiveness conditions; semantics and interpretation are not. The second remark, which is related, has to do with science and

[38]The U.S. National Security Agency.

naturalisation. In cognitive science and computer science, if not in logic or philosophy per se, it is common to think that the logicist tradition, in virtue of its commitments to formality and to the (at least potential) mechanisation of inference, is thereby rendered naturalistically palatable, in the following strict sense: that logical systems, *in virtue of being formal*, thereby somehow secure an at least potentially causal explanation.

Nothing could be further from the truth. Yes, as just suggested, there are some reasons to suppose that the upper half of figure 3—having to do with proof, syntax, and inference—may be amenable to causal explanation (though even showing that is going to take contentious reconstruction). But nothing in the picture laid out above provides any reason to suppose that the semantic interpretation relation I, or the semantic domain D—or indeed any interesting semantic property—need necessarily succumb to causal account. Logicians, in my experience (as opposed to logically-oriented computationalists), are *mathematicians*, not *naturalists*.

Indeed, some of the most prominent results in logic, such as the incompleteness theorems, could not even be formulated in purely naturalistic language. But we don't need any such radical conclusion here. For our purposes all that matters is that questions of what secures the interpretation function I, what sort of account semantics will ultimately be explained by, either in a particular case (such as arithmetic) or in general, is not required, by anything in the logicist framework, to be naturalised or even naturalisable. Perhaps semantics can be naturalised; perhaps cognitive science will show us how to naturalise semantics.[39] But *logic* doesn't show us how.

This negative observation will figure centrally in the upcoming reconstruction. As intimated above, various prominent counterproposals to the logicism—from theories of self-organising systems to proposals to understand cognition as a dynamical system to literary philosophies of the body—are, in this sense, more conservative than the logical tradition from which they aim to free

[39]Exactly this is the aim of such projects in philosophy of mind as Fodor's asymmetrical dependency theory, Dretske's informational account of semantics as counter-factual-supporting correlation, Millikan's theory of semantics as grounded in a biological notion of proper function, etc. «refs»

themselves, in virtue of being more committed (in advance) than logic to a scientifically traditional form of causal explanation.[40]

2e Norms

The third remark about the logicist framework summarized in figure 3 is very important. Logical systems are **normatively constrained**: strong evaluative metrics govern the ways in which the four ingredients (syntax, semantics, operations, semantic interpretation) are tied together. The importance of these norms cannot be underestimated; they are utterly critical to the stuff and substance of representational schemes. Without them, the whole apparatus of logic (and representation more generally) would collapse.

The primary norms embraced in the logical tradition are **soundness** and **completeness**: two versions of a rough requirement that what is derived (by $\mathcal{P}$) correspond to what is true (in $\mathcal{D}$). Though 'soundness' and 'completeness' are not very symmetric *terms*, the norms have a clear symmetry. Systems are sound just in everything that can be derived (formally, effectively) in the syntactic realm $\mathcal{S}$, from a starting set of premises, is true or valid,[41] in the semantic domain $\mathcal{D}$; systems are completeness just in case the converse is true: everything that is true or valid in the semantic domain $\mathcal{D}$ can be derived in the syntactic realm $\mathcal{S}$.[42] Normally, one simply proves or demonstrates the soundness of a system, and the shows its completeness (if things work out well[43]). But to prepare us for a more general account, we can think of this as a two stage process. First, soundness (truth-preservation) and completeness are specified to be the governing norms. Soundness and completeness, that is, should in the first instance be understood as *regulative*; then the proofs that the given system is sound (and perhaps complete) should be viewed as *demonstrations that the system in question has met its regulative constraints*.[44]

---

[40]«Highlight this irony»

[41]I am not distinguishing truth and validity here …

[42]A more general reading of soundness and completeness is given in §■■, below.

[43]«Put in a note about completeness, in model-theoretic guise, often being a sham»

[44]«Insert a sidebar on the division of labour between *truth* & *soundness*—

When logical system are presented, traditionally, the syntax, grammar, proof regimen, interpretation function, etc., are all usually simply laid out in ostension—as if they had arrived, full-blown, as "facts" for theoretic consideration. But lurking underneath this symmetric presentation is a critically decisive asymmetry: whereas, as we have already seen, it is the formal or inferential facts (i.e., issues having to do with the upper half of figure 3) that are subject to constraints of effectiveness, it is the semantic facts—interpretation, truth, validity, etc. (i.e., issues having to do with the lower half of the diagram) that, from a normative point of view, are in the driver's seat. That is: at the most general level, the normative constraints on a logical system take the following form:

> *The (upper-level) effective transitions are normatively regulated to honour the (lower-level) semantic facts.*

This general pattern—of the effective mandated to honour the semantic—is as deep a fact about logic as there is. It will stand with us throughout our upcoming reconstructions (indeed, it survives even much more radical reconstructions than we are able to assay here[45]). If all one wanted were a *causal construction kit*, one would be crazy to choose logic. What logic gives us is something radically more substantial: *normatively-governed* construction kits.[46] Without norms, logic would be an empty vessel, devoid of substance—uninterpreted mechanism flapping aimlessly in the breeze.

2f Independence

The fourth and final comment about logicism has to do with the relation between the syntactic and semantic realms implicit in the

---

i.e., between what parts of the "worth" of a logical system are supplied by the language and inference and interpretation rules, and what parts by the axiomatization of the task domain [[cf. the division of labour between the calculus and the laws of motion in physics]].»

[45]Smith, Brian Cantwell, *On the Origin of Objects*, Cambridge: MIT Press, 1996.

[46]That's not quite fair, of course; what logic gives you—as I hope to make clear before the end of the paper—is a radically specific form of *semantically governed* construction kit.

picture we have been working with (again, see figure 3). In particular, the overall picture is constituted against three independence and one dependence claim.

The first independence claim concerns the two basic realms in terms of which logic's conceptual framework is articulated. Paradigmatically, the realms are established—or, as one typically but curiously says, "specified"—independently: one delineates them in separate stages, giving each its own autonomous ingredients, identity conditions, etc. Since modern logic was developed to deal with issues of mathematical inference, it may be that the ontological character of the realms was assumed to follow from whatever ontological conditions warrant the metaphysical existence of general mathematical entities. But whatever the reason, the two realms are assumed to be autonomously specifiable.

The second independence claim (again separating the two realms) is implicate in the so-called **formality condition**, mentioned earlier—the universally-accepted requirement that the operations or transitions on S constitutive of P be defined purely in virtue of the (positive) form or syntax of the elements of S, *independent of those expressions' semantic interpretation* $I(S) \in D$. The formality condition, that is, is s second way in which the realms S and D are separated.

Third, the interpretation function I is critically assumed to exist independent of (and again explanatorily prior to) the operations or transitions constitutive of P. This is essential in order for the governing norms to take the form that they do. It would be ill-formed (i.e., would violate this third explanatory autonomy) to take a sentence S to mean something like "This very sentence has not yet been derived"—since in such a case S could be true, *but could not be (soundly) derived*. In general, the conceptual structure of soundness and completeness requires that the interpretation be established (or exist) *prior to and independent of* the operations, in order for the normative constraints on the operations to honour it. If A's job is to honour B, then B had better be defined independent of A, or else one runs the danger of setting up a vicious cyclicity.[47]

[47]This way of putting things exaggerates necessity, though not the accepted structure of formal logics. As we will see, it may be possible to defined

But then, once the realms are all separated out in this way, and these strong independence standards are in place—i.e., once the autonomy and separateness of the two realms is firmly established in all the requisite ways—then the norms operate *exactly by tying the two realms back together again*. It is this **reconciling tug**, as I've said, that gives logical systems "bite." In a way, the underlying conceptual structure is almost ironic: first one ensures that everything is cleanly and totally and utterly kept apart (logically, conceptually, ontologically, whatever) so that, once things are separated, they can be *regulatively brought back together again*.

It is going to be of the utmost importance to determine what the initial separateness, and what the subsequent tying back together again, come to in an adequate representational reconstruction that is suitable for embodied cognition. For now, it is enough to see that the very raison-d'être of a logical system derives from this never entirely reconciled but nevertheless reconciling tug between the two realms. Minus semantic interpretation and governing norms—i.e., as a pure structural construction kit—logical and representational systems are wimpy. For purposes of sheer assembly, abstract Erector sets, hydraulics, or C++ would be vastly better—or even, for that matter, carbon-based molecules or DNA.

(semantic) interpretation *partially* in terms of operations, without rendering the resultant norm vacuous. But constitutive interdependence of this sort is one of the radical generalisations we will take up in §■■; it is never, so far as I know, employed in a logical system.

3 Reconstruction I • Computation

It is difficult to say exactly what it is about the classical picture that troubles proponents of an embodied approach. But at least eight properties have drawn comment from various writers. They are listed in figure 5, with the presumptive character of logicist models indicated on the left, and the properties recommended for a new, embodied or situated conception of cognition on the right. I make no claim that this list is complete, that the issues it enumerates are independent, that it does justice to all anti-classicist sentiments, or that it is correct in its characterization of logic (in fact I will presently argue that at least one entry in the left column is false). But it will serve for our purposes.

Crucially, the list doesn't mention *representation*. It is critical to our project, however, to recognize that one of the arguments frequently heard in the "embodied cognition" camp is that it is exactly in virtue of being representational that logic exemplifies the properties identified on the left—and therefore that, in order to manifest the properties listed on the right, a system must abandon representation.

For an advocate of generalisation, therefore, who resists (especially in *a priori* form) this strong antirepresentationalist stance, the tabulation raises two challenges: (i) to understand which of the characteristics in the left list are true of only a particular (logicist) species of representation, rather than of representation in general; and (ii) concomitantly, to the extent that any of the characteristics listed in the left-hand column turn out in fact to be species-specific, to understand how a generalised conception of representation can deal with the corresponding property identified on the right.

Given our concern with reconstruction, however, we first need to analyse the generalist's starting assumption: whether the characteristics listed on the left hand side of the table really do hold of representation on a logicist conception. That is: to what extent is the left-hand column *correct*?

	Logicist	Issue	Embodied
1	*Abstract, disembodied*	Materiality	*Concrete, embodied*
2	*Explicit, linguistic*	Vehicles	*Tacit, non-representational*
3	*Disconnected*	Environment	*Fully engaged*
4	*Separate, independent*	Realms	*Not separated*
5	*Static, atemporal*	Temporality	*Dynamic*
6	*Digital, discrete*	Character	*Continuous*
7	*Context-independent*	Interpretation	*Context-dependent*
8	*Ratiocination, thought*	Activity	*Improvisation, navigation*

Figure 5 — Dimensions of Differentiation

Start with the first point of alleged difference between logicist and embodied views: the claim that classical logic treats its subject matter abstractly—and thereby fails as a model of human cognition, because of its consequent inability to deal with important facts about humans' material embodiment.

At least in its first half, regarding the abstract treatment of formal logic, the claim seems true on the face of it. Not only does model theory almost universally analyse semantic realms in terms of purely abstract set-theoretic domains,[48] but even syntactic realms, while somehow vaguely concrete (for example in the sense of sustaining an idea of syntactic *tokens*, and being subject to mechanical realisability) are still not treated in reigning theories *as* concrete—as bluntly physical or material in any important (e.g., energetic) sense.

As already intimated, however, I believe that although this abstract view is socio-intellectually or epistemically correct about how logicians treat or analyse logic, it is *ontologically* misleading. It

[48]By 'purely' abstract I mean a set all of whose members (recursively) are abstract. In contrast, suppose A is a two-element set containing elements B and C, where B and C are also sets—B a set of camels and C a set of zebras. In such a case all three sets, *qua* sets, may arguably be considered abstract—but since the inner two are made up of concrete elements, I would not consider any one of the three *purely* abstract.

conveys the idea that logic is constituted abstractly, without impact or constraint deriving from physical reality. Surface appearances notwithstanding, and pace the protestations of practicing logicians, I will argue that the entire substance of the traditional logicist view rests on very real constraints that derive directly from a logical system's concrete materiality.

To see this, though, we need to step back from logic for a moment, and approach the subject matter of representation and computation from a far more general perspective than usual. That will be the task of this section; we will return to logic and the logicist model in §4.

3a Meaning and mechanism

The most fundamental issue underwriting representational and computational systems—and, more specifically, the issue that underwrites the classic logicist tradition in cognitive science in particular—is the interplay between **meaning** and **mechanism**. So important is this issue, this contrast, this generative tension, that in other writings I have dubbed it the **primary dialectic** of the intentional sciences. What it comes to depends on what one takes 'meaning' and 'mechanism' to mean; but at a very rough level, the question is something like the following:

> How can things that are entirely concrete—no magic, spirits, divine intervention, etc.—*without violating that inexorable underlying materiality*, nevertheless, in the appropriate sense, "transcend" that materiality, so as to think, dream, mean, wonder, refer, be right and be wrong?

I have called this a dialectic, but that does not mean it is an outright opposition. Cartesian predilection notwithstanding, few believe that meaning and matter are opposites or distinct substances, in the sense that the world consists of those two kinds of things, glued together with God's own epoxy of set theory. Rather, at least for materialists or physicalists, the question is how ordinary bodies or mechanisms, which in one sense *are* merely physical, in another sense are *not* merely physical, but must instead authentically and legitimately be understood (perhaps even constituted) in intentional terms?

I believe it is impossible to understand the whole edifice of

syntax, semantics, formality, truth, soundness, etc., as adumbrated in the previous section and refined over more a century of academic scholarship, except as an attempt to instantiate a plausible answer to this daunting metaphysical question, albeit in an extraordinarily restricted setting.

3b Effectiveness

It would be natural to assume, of this dialectical pair of meaning and mechanism, that the meaning or semantics side (truth, meaning, representation, content, etc.) would be the troublesome element. It would be natural to assume, that is, given 300 years of spectacularly successful natural science, that we would have an adequate and even good grasp on the material or mechanism side.

It would be natural—but it would be wrong. It turns out that coming to grips with the "mechanism" half of the dialectic has proved almost as difficult as understanding meaning and truth.

The notion that has been in primary focus, in the quest to tame the mechanical, has been that of **effectiveness**—as betrayed in the fact (already mentioned) that the reigning mathematical foundational theory taught in computer science departments is called the *theory of effective computability*. As it happens, I have grave doubts as to whether this vaunted theory merits its ubiquitous name "theory of *computation*," but that it focuses on effectiveness is surely right. The aim behind this body of work has been to formulate, in as clear and theoretically profound a way as possible, what can be done, by a concrete physical mechanism—both absolutely (i.e., without restriction on time, space, or other finite resource), and relatively (in the sense of with relation to more or realistic constraints on allowable resources bounds).

These issues have been explored in what seem to be relatively abstract systems, under the guise of syntax, proof theory, and numerical computability. Theoretical results are by and large framed mathematically (e.g., in the difficulty of solving this or that mathematical problem, or the complexity of, for example, factoring products of large primes). It is this rampant mathematicization that, I believe, though not problematic on its own, has within the larger scheme of things proved radically misleading. In another place I argue at length that all computability results—both absolute, as in Gödel's incompleteness results, the unsolv-

ability of the halting problem, etc., and relative, as in the results of complexity theory, the difficulty of deciding classes of formulae, etc.—derive directly from physical, material constraints on underlying mechanisms. Sure enough, in the theory as we know it the results are *framed* mathematically, but so are (at least many) results in physics and chemistry. Present theoretic practice notwithstanding, the subject matter of theoretical computer science is by my lights entirely concrete.

Let me admit straight up that this is a contentious claim; I have yet to meet a logician who believes it. Informally, though, in my experience, most working computer scientists not only *believe* it, but so thoroughly *assume* it that it takes work to show them that it is not in fact something that logicians presuppose.

What makes the issues subtle—but at the same time interesting—is that it is clear that major computability results are not specific to any *particular* material substrate. Factoring primes is approximately equally difficult, whether one uses vacuum tubes, silicon transistors, or even tinker toys. This betrays what I dub the **secondary dialectic** underlying computing: between the *abstract* and the *concrete*. My claim is that, whereas physics (and perhaps material science) has focused on the completely concrete, and mathematics (presumably) on the completely abstract, the "natural home" of computability results lies somewhere in between—*but much closer to the concrete end than is normally (especially theoretically) realized*.

Historically, the reasons why the formal "theory of computability" has framed its results mathematically are sure many, including (but not limited to) the following:

1. It is a perfectly evident observation that computation, at the level at which theories have dealt with it, can be, as it is said, *multiply realised* on a wide (perhaps even limitless) variety of different substrates;
2. The theoretical aim has been to identify very general results, rather than specific material concerns (for example, it is only recently that computer science has begun to deal with real-time results);

3. Historically, the tradition developed out of concerns with metamathematics, making an abstract perspective more natural;
4. Scientific results are almost always expressed mathematically; since the computability results were not framed in terms of any readily-identifiable units (kilograms, ergs, etc.), the equations appear to traffic in purely numeric quantities;
5. Since the problems for which computability results were developed had primarily to do with mathematical subject matters, the languages used to represent them were by and large context-independent, which turns out to imply that one could frame results purely in terms of types, without regard to concrete specific facts about individual tokens (the way one needs to do when treating indexical expressions, for example).

Of these five, the first (multiple realisability) is indubitably most famous, but the last, having to do with the relation between types and tokens, may cut the deepest. Since the subjects matters taken up in the theoretical context have (contingently) been primarily abstract, it has proved convenient to deal with them abstractly. But what it is that is abstract, in my view, has been misinterpreted. In particular, I argue:

> Reigning theory of logic and computing treat computational entities (states, marks, etc.) as *abstract individuals*, whereas in fact they are more properly understood as *concrete type*—i.e., as types of concrete things.

The reason why the difference matters is because the constraints on the notions (what it is to be a state, what it is to be a mark, where the properly-vaunted computability come from, metaphysically) derive from the concrete, physical world—the world of which they are types, rather than from the abstract, logical, or mathematical world (where types presumptively "live").

Arguments supporting the changer in perspective rest on such facts as that, if one changes the physics of the realizing substrate, one can change complexity results at will. Intimations of this were

recognised as early as in the 1930s by Robin Gandy,[49] who showed that the absolute computability results depended in immediate and subtle ways on the character of the physical mechanisms on which they were assumed to be implemented.

In the end, though, the proof of any theoretical claim rests heavily on its theoretical utility. Some of the arguments I advance are negative: that not recognising and understanding the physical nature of effectiveness leads directly to various negative entailments: one can solve the halting problem, one cannot explain the ubiquitous notion of a "reasonable encoding," etc. The lion's share of the argument, however, rests on positive results: that if one does recognise the concrete nature of effectiveness, one can (among other things) achieve the following sorts of results:

1. Explain the notion of a reasonable encoding (both what the constraints on being a reasonable encoding are, and also why the notion of a reasonable encoding has received so little theoretical attention);
2. Make sense of the rise of Girard's linear logic, computer science's interest in intuitionistic type theory and constructive mathematics, etc.[50]
3. Predict the proposed fusion of foundational theories of quantum mechanics and computer science-based theories of information;
4. Make sense of why physicists are interested in super-Turing computability, continuous models of computation, quantum computing, etc.; and
5. Resolve otherwise unexplicated tensions between what is real and what is virtual (e.g., in popular conceptions of computational technology).

In spite of these benefits, the proposed adjustment in our understanding is not without cost. For example, it is a inescapable consequence of reconstructing the current (so-called) theory of effec-

[49]Gandy, R. (1978), 'Church's Thesis and principles for mechanisms', in K. J. Barwise, H. J. Keisler, and K. Kunen, eds., *The Kleene Symposium*, Vol. 101 of *Studies in Logic and Foundations of Mathematics*, New York: North-Holland, pp. 123–148.

[50]«Ref Girard, Martin-Löf, etc.»

tive computability as a theory of *effectiveness* that it emerges from that reconstruction as no longer being a theory of *computing*, because it deals with only the first (mechanism) arm of the primary dialectic, not with the second (meaning, semantics, reference, truth, etc.). When conjoined with the present point, that the underlying constraints that give substance to the theory are direct consequences of the concrete, physical nature of the underlying medium, one is forced to conclude that what is universally known as the theory of effective computability is, in point of fact, (and presumably will eventually be historically recognised as) a **mathematical theory of causality**—namely, a theory of what can be done, in what time and with what resources, by what sorts of arrangements of concrete, physical stuff. That such a theory should be framed at some level of abstractness, away from very specific concerns having to do with particular materials, is entirely to be expected. It is for this reason that I have dubbed the properties that the theory traffics in *effective* properties, rather than *physical* properties; they are properties that systems (or states) can *do consequential work in virtue of possessing*.[51]

Two final remarks.

First, a proponent of embodied cognition might argue that even if we do reconstruct computability theory as a theory of causality, it will still be too abstract for cognitive science: that in order to understand cognition "in-the-wild,"[52] one needs to understand not only relatively abstract causal properties of the system, but quite concrete properties (such as heft and materials)—e.g., in order to understand rhythm and dynamic movement. That may be, but there is every indication in theoretical computer science that the theory in question is rapidly being refined so as to deal with more and more direct physical parameters (in order, among other reasons, to treat issues of three-dimensional packag-

[51]It is also unclear exactly what it is to be a physical property. Being a million light-years from Alpha Centauri is presumably a physical property, but not an effective one; it would be impossible, at least in any remotely practical sense, to build a device that could "detect" the exemplification of this property.

[52]The term is from Edwin Hutchins, *Cognition in the Wild*, Cambridge: MIT press, 1996.

ing and real-time computing). Moreover, the embodied cognition movement has to be interested in bodies and materials at *some* level of abstraction. Suppose one were to replace the control circuit for the muscles of an animal with an electronic souped-up version; what matters, presumably, even to the most materially-oriented theorist, is that the signals match, that power be supplied, that the right function be computed in real-time, etc. There are questions of whether such implants could work—and how much of our cognitive facilities could be upgraded in this way. But virtually no one thinks that a brain implant would literally have to be made of DNA-based neurons, in order to function in a "materially" appropriate way. Put it this way: neurophysiology and the theory of effective computability are climbing up the same mountain, even if from different sides.

Second, let me reiterate what I hope is clear: that reformulating our understanding of (what is known as) computability in concrete, material terms, in the recommended fashion, is an enormous as-yet open intellectual task. My guess is that it will take decades for the transformation to take place. For example, all absolute and relative computability results—that whether a Turing machine will halt on an arbitrary input cannot in general be algorithmically decided, that factoring primes is hard, etc.—will have to be reformulated as issues about mechanisms, not issues about *numbers* or *decisions*.

Nevertheless, this first reconstruction of computation—recognising the physical character of the notion of effectiveness that constitutes half of the primary dialectic on which computing rests, and that serves a lynchpin in our understanding of logicism—is a necessary prerequisite, I believe, of understanding the essential character of representation.

4 Reconstruction II • Semantics

Turn then to the second arm of computing's primary dialectic: meaning and semantics. There is a major reconstructive move to be made on this side, as well, again have to do with physicality. This time, however, the issue is not with the relation between the abstract and the concrete—what I called the secondary dialectic. Indeed, in order to get at it, we first have to set a potentially distracting of abstractness aside.

4a Models

Consider semantic domain *D*.[53] As we saw, in classical logicism this realm is usually treated abstractly: as a set-theoretic construct of objects, properties, and relations, perhaps extended with functions, situations, states of affairs, facts, propositions—and sometimes possible worlds. It is not that the atomic objects on which this construction is based need necessarily be abstract—i.e., it is not that *D* need necessarily be purely abstract[54]— but rather that the composite structure into which the objects and properties and such are assembled, for semantical purposes, is (again, typically) more of a mathematical structure than it is, say, the full disheveled situation out the window.

One self-evident generalising step already presents itself, therefore: if the embodied cognition movement aims to deal with material creatures interacting with their environments, we will have to adjust our conception of semantics so that semantic domains don't just include concrete individual objects "at the bottom," as it were, but are themselves *full concrete environments*, such as train station platforms in modern Tokyo, or the messy situation where the Amazon pours out of Brazil into the Atlantic Ocean.

There is a methodological subtlety here. The reason that semantic domains are paradigmatically mathematical or abstract is that, in the classic tradition, semantics is usually studied **model-theoretically**. The semantic interpretations of representational vehicles are analysed in terms of abstract (set-theoretic) models or "stand-ins" for what I will call the genuine target domain, rather

[53]Figure 3, p. ■■
[54]See fn. ■■ on page ■■.

than directly, in terms of the real target domains that the vehicles are authentically about. The situation is depicted in figure 6. Suppose we construct a logical axiomatisation of the patterns of car movements on the expressways surrounding New York. D, in the figure, would be the actual, metaphysically occurrent concrete situation on the roads around town; M would be a *mathematical model* of those freeways, in terms of which the formal semantics would be formulated.

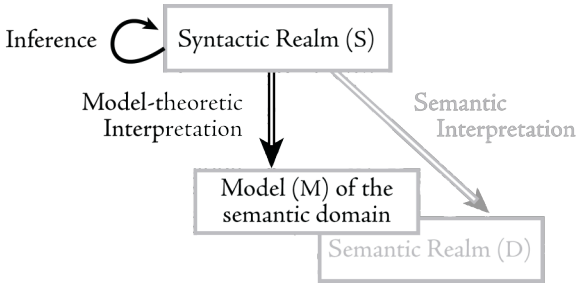

Figure 6 — Model-Theoretic Semantics

The rhetorical situation in this situation is complicated by the (often-noted) fact that the term 'model' is used technically, within the logical tradition, in a non-standard way. It in technical or theory-internal contexts, logicians speak of an element or structure of M's being a *model of a sentence S* (or of some other syntactic or representational entity). But that way of putting things splits from lay parlance, which would more likely call *M* a model of the target domain *D*. Thus imagine some aeronautical engineers producing a blueprint of a new kind of airplane wing, designed (say) to avoid turbulence at high lift. Suppose, to test the design, they build a plastic model to try out in a wind tunnel. Normal parlance would call the plastic device a *model of the wing*, not a *model of the blueprint* (i.e., would call *M* a model of *D*, not a model of *S*). In logic, however, the entity that is analogous to *M* is said to be a *model of the sentences*—i.e., a model of the entity that is analogous to the blueprint.

On the face of it, this is just a terminological ambiguity, so confusion need not reign so long as we keep usage clear. (In this paper, since my audience is cognitive scientists, not logicians, I will side with lay practice, and talk about *M*'s being a *model of the target domain*, not of the representational vehicles.) But more serious issues arise if, forgetting that *M* is a model, we mistakenly take (what I will call) **insignificant** properties of *M*—i.e., proper-

ties of *M* that are *not* intended to model anything in *D*—to be part of the interpretation of *S*.

It is a truism, after all, that not all properties of a model *M* can be intended to represent or model properties of that which it models. Thus in our example, the model of the airplane wing was made of plastic, but presumably with no implication that the aircraft wing was to be fabricated of plastic. Similarly, the cost of the model presumably bore no modelling relation to the cost—or indeed to any other property—of the thereby-modeled wing. Of course one can construct examples in which these things are false: one could construct a wing in which materiality (plastic, wood, whatever) of the model corresponded, directly or indirectly, with some property (perhaps the materiality) of the thereby-modelled wing. The point is only the following: only some (usually a finite number) of the (infinite) properties of a model are ever intended to correspond to only some of the (infinite) properties of what is modelled.[55]

For our present purposes, what matters about the model-theoretic approach to the semantics of logic has to do with its abstractness. In particular: *From the abstractness of set-theoretic models, nothing* (necessarily) *follows about the concreteness or abstractness of genuine (target) semantic realms*. Methodological abstractness, that is, need not vitiate subject matter concreteness. So discussions of the issue of abstractness vs. concreteness—item number 1 in figure 5 (page ■■)—should not be influenced by the fact that logicists do semantics model-theoretically. Doing so is compatible with arbitrarily concrete commitments about the nature of the semantic domain.

At the same time, we mustn't assume that it is intrinsic to embodied cognition that the relevant semantic or task domains must

[55]What is really going on here is that the relation between *M* and *D*—what I am calling a "modelling" relation—is itself a semantical or representation relation, and should be studied as such. In part because of this use of semantical relations to study semantical relations, and also because of the point raised in the text—that it is tempting to forget which properties of the model are genuinely significant, and then inadvertently to take insignificant (non-modeling) properties of *M* to be significant—my own preference is to avoid model-theoretic semantics entirely, in favour of what I would call **direct semantics**.

be entirely concrete (i.e., talked about in terms of bare materials). Suppose an agent is designed to proceed in the face of a single obstacle on its path, confident that its navigational skills will allow it to negotiate its way around one thing, in real-time. When it encounters a group of more than one obstacle at once, however, it is designed to stop and plan a deliberate route around them. Are "groups of obstacles," or "routes," concrete or abstract? Who knows? And no matter how concrete the domain into which an embodied agent wanders, it will always be true (minimally, because of finite resource bounds) that such agents will need to deal with those domains at some level of abstraction.

Put it this way:

> *For cognitive science to deal reasonably with embodied and embedded cognitive agents, the secondary dialectic adumbrated above, having to do with the relation between the abstract and the concrete, will be as applicable to environments and task domains as it is to creatures and cognition itself.*

4b Formality

Setting issues of abstractness provisionally aside, then, turn to the second critical reconstruction, this time having to do with semantics and formality.

It is a prominent and profound fact about logicism that logical systems are considered *formal* systems; that logic is the product of the *formal* tradition, that to construct a logical model of something is often identified with *formalising* it. Just what 'formal' means, however, is one of the most diabolically complex issues in this entire subject matter.[56]

Overall, there are two rough sense of formal that need to be distinguished. The first, which I will call *methodological*, has to do with what it is for logic (and perhaps computing) to be a formal discipline, what it is to "formalise," and the like. I will not deal with these concerns here except to say that they seem intimately tied up with expectations and assumptions mentioned earlier: about naturalisation, about the possibility of giving explanatory scientific accounts, about the possibility of mathematical analysis, and the like.

[56]«Ref AOS»

What I do want to focus on is another set of formality intuitions, this time more *ontological* in character, having to do with how such systems are as a matter of fact structured, with how they work. Some of these intuitions were mentioned earlier, in §2. In particular, it is taken to be a criterial condition that inference (proof, operations) work or proceed *formally*. This requirement, which, as mentioned in the previous section, Fodor has dubbed the **formality condition**,[57] is viewed as an absolute mainstay of the classical representational tradition. It is thought to bring to logic and computation, and thereby to cognitive science, its strongest weapon in the struggle to resolve the primary dialectic, and thereby to finally defeat the threat of the mind/body problem. Indeed, there are those who would say that formality is the very foundation on which the material success of the classic tradition relies.

It was also mentioned earlier that the formality condition has two different senses. The *positive* aspect of formality has to do with shape, syntax, grammar, or "form"; it militates that inference operations be definable (and work, causally) in virtue of the syntax or form of the constituent expressions (representational vehicles). It is the *negative* aspect of formality, however, that concerns us here: the ubiquitous assumption that both the syntactic properties and identity conditions on the expressions or representational vehicles (elements of *S*), and the operations or effective transitions defined over them, must be defined *independently of semantics*.

It may be that one of the *consequences* of the negative reading of formality has to do with naturalisation: that some of the overall logicist story (at least the upper half of figure 3) will be amenable to causal account. The original *motivation* underlying the negative reading, however, stems from a very basic insight about representation in general. And that is the insight we are after. For one the most serious Achilles' heels in the embodied cognition stance, as suggested in the introduction, is that, in distancing itself from the formal logicist tradition, it runs the risk of missing this insight, that underwrites formality: an insight that not only implicitly undergirds the classical tradition, but that cuts to the very heart of

[57]See the sidebar on p. ■■.

what representation is like—indeed, to what representation is for.

The idea is this. Genuine semantic properties—being true, referring to Cheops, etc.—are not of the right sort to figure in how symbolic or representational systems actually work. *Semantical properties*, to put this in terms we have already used, at least in general, are *not effective*.

Intuitively, one reason semantical properties aren't effective is that they are often (perhaps always) *relational*. The truth of the sentence "dinosaurs were warm-blooded" seemingly depends on facts that obtained hundreds of millions of years ago—facts that, in a rough and ready sense, are simply *too far away* to do any work in affecting the right-here, right-now microdynamics of how an inference system works (human or machine). More generally, representations often bear semantic relations to situations or states of affairs that are *distal*, and distal things, because of the locality requirements of physics, simply cannot get into the act in affecting the here-and-now.

There is a discrepancy, that is, between:

1. The local, effective, immediate structure of a representational system, in terms of which it (causally) works; and
2. The paradigmatically distal, non-effective, semantic structure of the system, in terms of which it is normatively characterised.

The dialectic is mortal. Nothing will matter more to the story to come than the interplay between these two kinds of property. Indeed, it was already evident in the normative structure of the classical model we started with that what representational norms *govern* is the syntactic or proof-theoretic or effective local workings of the system, whereas what the norms are based on or are designed to *honour* are the system's semantic contents. We will get back to norms presently; what we can do here is to state, very simply, what I will take to be our second reconstruction—a reconstruction of logicism's commitment to a (negative, ontological) reading of formality:

Semantic properties aren't effective

Semantic properties, that is—the "orienting towards the world" properties in virtue of which representational systems are norma-

tively governed— cannot in general be assumed to be effective, in exactly the sense of "effective" we talked about in the first reconstruction, above. They are not properties in virtue of the possession of which systems can make concrete things happen.

Three comments.

First, there is a scope ambiguity in the foregoing statement: whether it is being claimed that *no* semantic properties are effective, or only that *not all* semantic properties are effective (i.e., that one cannot conclude, in virtue of a property's being a semantic property, that it is thereby guaranteed to be an effective property). Call these the **strong** and **weak** readings of the reconstruction of formality, respectively. As will emerge later on, I believe that the strong reading is true; but for now we can make do with either.

Second, on the (negative) ontological reading of formality within logic itself, the claim was made that formal systems operate *independent of* the semantics of their ingredient states. "Independence" turns out to be a notion not unlike modality; it comes in strengths: logical independence, ontological independence, metaphysical independence, etc. It is no aim of mine here to say which notion of independence the formal tradition is committed to. What we can see, however, is that an independence claim is stronger than the "semantics is not effective" version just formulated (thus we are already starting our second strategy, of generalising).

In particular, we are in a position to begin to see what is wrong, or anyway too restrictive, about classical formality. Formal logic essentially infers, from the (manifest) *non-effectiveness* of the semantic, that the workings of the system must be *independent* of semantics. Sure enough, if semantics is not effective, then how a system works cannot depend on semantics in any very full (at least in any causal) way. But—and this is a critical generalising point—*there is a world of difference between non-dependence and independence*. That this is true is made obvious by reflecting on human affairs: someone can take your views into consideration, in forming their opinion, without adopting either extreme: of being slavishly dependent on what you think, or being so independent of what you think as to be wholly autonomous and uncaring. In

human affairs, both limits are recognised forms of pathology. Between the two lies an entire realm of *partial dependence*—or perhaps better described, *partial interdependence*. To foreshadow a bit, partial interdependence, of this rough sort, will eventually emerge as the constitutive relation between (reconstructed) syntax and semantics—that is, between concrete, "make-it-happen" effectiveness, on the one hand, and non-causal directed-to-the-world normative governance, on the other.

But we are getting ahead of our ourselves. The point is that from an ontologically point of view, formality is wrong, because too extreme. But it rests on a profound insight, about the non-effectiveness of the (normatively-governing) semantic. Preserving this insight, and understanding its import in cases of embodied, embedded, engaged cognition, is the key to the challenge identified in the opening sections: understanding how to retain semantics through a transition from the abstract to the concrete.

5 Towards a participatory account

Before we turn in full force to the second, generalisation phase, it will help to summarise how far we have come. For already the outlines of a more powerful conception of representation can be discerned.

5a Logic, formality, and concreteness

What we are driving towards is a profound dialectical interplay between the effective and the non-effective. At the deepest level, I claim, this dialectic (albeit in a restricted form) has underwritten—has always been what matters most—about the logicist program. All sorts of familiar (and essential) features of the logical conception can be reconstructed in its terms. If we take "**rehabilitation**" to mean "reconstruction plus generalisation," then:

1. The "effective" structure of a representational system is the rehabilitation of *syntax* or *form*;
2. What the system does, mechanically, is the rehabilitation of *proof theory or inference*;
3. The situations or states of affairs in the world towards which the system is (normatively) oriented is the rehabilitation of *semantic interpretation*; and
4. The fact that the system is not (in general) effectively coupled with those situations towards which it is normatively oriented is the rehabilitation of the claim that inference operates *independent of semantics*.

The last of these claims of course has to do formality. Throughout the discussion so far, I have identified two different (ontological) readings of formality: a positive reading, having to do with syntax or "shape" or "form," and a negative one, meaning "independent of semantics." As should by now be evident, our two reconstructions dealt with the positive and negative readings, respectively:

5. The positive reading was reconstructed in terms of the fact that systems work in virtue of the presence of effective (concrete, causal) properties;
6. The negative reading was reconstructed in terms of the absence of effective coupling with the semantic domain.

The Representational Mandate

1. Conditions
   a. A representational system must work, physically, in virtue of its concrete material embodiment (the role of effectiveness).
   b. But it is normatively directed or oriented towards what is non-effective—paradigmatically including what is physically distal.
   c. Being neither oracle nor angel, it has no magic (non-causal, divine) access to those non-effective situations; just caring about them is not enough (physical limitations bite hard!);
2. So what does the system do?
3. It
   a. Exploits local, effective properties that it can use, but doesn't (intrinsically) care about—i.e., inner states of its body and physical make-up, in interaction with the accessible (effective) physical aspects of its environment
   b. To "stand in for" or "serve in place of" effective connection with states that it is not (and cannot be) effectively coupled to
   c. So as to lead it to behave appropriately towards those remote or distal or other non-effective situations that it does care about, but cannot use.

Figure 7 — Representational Mandate

That is, the two reconstructions can be viewed as "**concretisations**" of formality—as reformulations in concrete, physical terms of something that classic logic has dealt with in an (unfortunately) abstract way. That concretisation will stand us in good stead with respect to the goal identified in §1: of preserving an understanding of semantics through a shift from the abstract to the concrete.

5b The representational mandate

In a sense, the moral so far is a recognition that concerns of concrete materiality and have lain submerged, just below the surface, in the traditional logicist conception—out of explicit theoretical

view, but nevertheless playing a critical role. What differentiates the new view is that those concerns are being brought into clear and unambiguous focus. In fact they almost define the character of the new view. For notice how thoroughly issues of concrete materiality permeate the emerging conception.

Representational or intentional systems, as we have seen, (at least typically) stand in semantic relations to distal and other non-effective situations. Such systems are normatively governed by those relations that they bear to those situations or states of affairs. But in a mechanical sense (on pain of violating physicalism), such systems cannot *work*, causally, in virtue of those relations—exactly because they are not effective. They can't work that way because the (semantic) properties tying them to those situations, and the properties of the situations that they are thereby tied to, are in general relational. So what do such systems do? They are constituted or arranged in such a way that they can use the (local) effective properties of their local, immediate structure—i.e., they use what is available to them: the effective properties of their causal ingredients, in conjunction with the effective (causal) properties of the environments in which they are deployed—*so as to behave, appropriately with respect to those distal and other non-effective situations.*

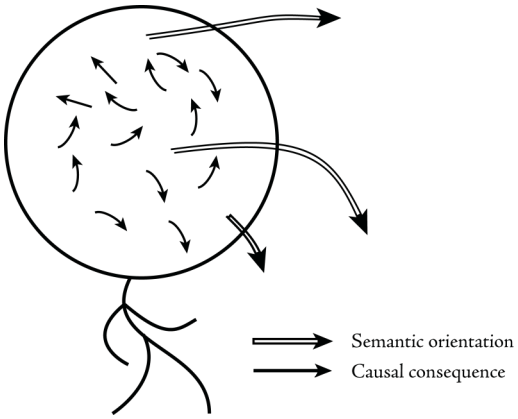

Figure 8 — Participation, First Pass

That is, a representational system:

> Exploits the effective properties of its inner states—properties that it can use, but doesn't intrinsically care about—to "stand in for," or "serve in place of" effective connection with states that it is not effectively coupled to, so as to lead it to behave appropriately towards those remote or distal situations —situations that it *does* care about, but that it *can't* use.

Or more simply yet, representational systems:

1. Exploit what is local and effective
2. So as to behave appropriately with respect to (to satisfy governing semantic norms regarding) what is distal and non-effective.

We still have to a considerable amount of work to do in order to see what this characterisation comes to in detail. But it will stand us in sufficiently good stead, over the long haul, to be worth a name. As indicated in figure 7 on page ■■, I will call it the **representational mandate**.

5c Coordination Conditions

A caricature of the view we are closing in on is given in figure 8. The system is constituted of a variety of states, and embedded in an (also causal) environment. In general, those states will exhibit two kinds of property:

1. **Causal consequences**, due to their effective properties, including the role they play in the overall machinery of the system (depicted as single-tailed arrows: '→'); and
2. **Semantic relations**, towards the states of affairs in the world to which the system is normatively oriented (depicted as double-tailed arrows: '⇒').

The stuff and substance of the system derives from the interplay between and among these two kinds of relation.

But figure 8 is too simplistic. It immediately needs to be generalised. First, it is not just the agent that is made up out of dynamic, efficacious states; the same is (in

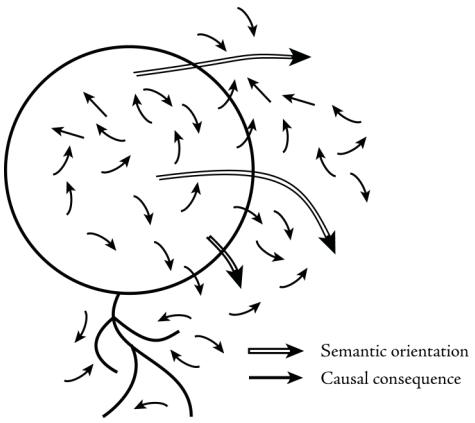

Figure 9 — Participation, Second Pass

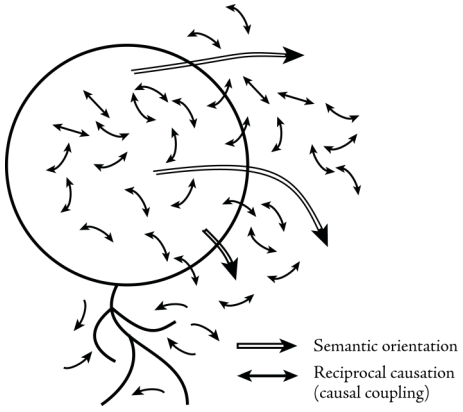

Figure 10 — Participation, Third Pass

general) true of its environment. So causal arrows need to be introduced into the environment. Moreover—in order to make room for the critical causal or effective engagement of the agent with the environment—arrows must also be added that cross the boundary (in both directions) between agent and environment. This much is shown in Figure 9. In addition, given that we are aiming at a general account (and with a nod to Newton's first law of motion), it is more general to change the (single-tailed) causal arrows to bidirectional ones, so as to license reciprocal causation, as indicated in figure 10. Finally, as shown in figure 11, it helps to indicate that semantic relations ('⇒') have vastly greater reach than causal arrows. They are not limited to states of affairs to which the system has effective access, but can leap across gaps in time, space, and even possibility, in dizzying array.

That is not to say that we have explained how arrows of semantic directedness are established, or even (metaphysically) what they are. Given a background physicalist metaphysics, they are going to depend on large-scale (distal and social) relational patterns, rather than on immediate patterns of local, effective coupling. But what is crucial to recognise here is that, once the two have parted company (for whatever ontogenetic reason), *it is the gap between them that al-*

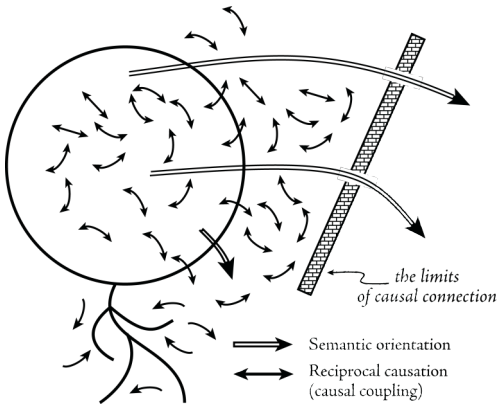

Figure 11 — Participation, Fourth Pass

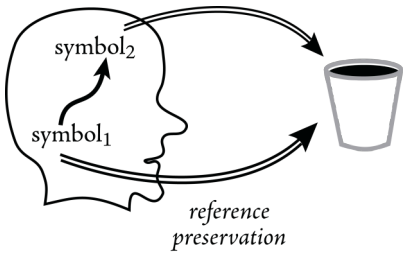

Figure 12 — Reference Preservation

lows normativity to establish a governing foothold. In fact such norms will eventually be identified as topological constraints on the relations between and among these two kinds of relations.

Simplistic caricatures of some familiar norms are shown in (structural coupling) the next set of figures. Truth- or reference-preservation—the traditional norm on sentential inference, and on term rewriting—is schematized in figure 12. Figure 13 depicts a basic constraint on perception: that a system, upon encounter with a situation ϕ, end up in (or construct) an inner state that represents ϕ. Similarly, a baseline condition on effectors is diagrammed in figure 14: that they cause to come into existence that situation that is represented by a state that triggers them.

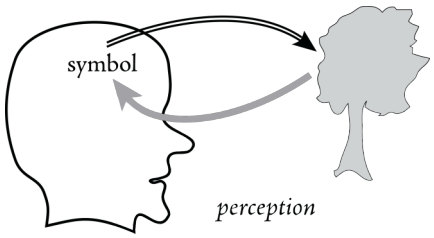

Figure 13 — Perception

Although it is reassuring to retain traditional norms, as soon as one applies this general framework to real-world systems,[58] it becomes evident that these are just three of a large number of potential normative constraints, some radically complex, and some of considerable interest to cognitive science.

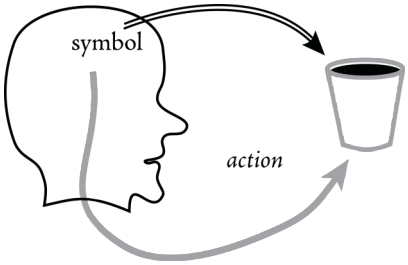

Figure 14 — Action

In the 1980s, when first working on these issues, I proposed a general framework in terms of which to analyse such norms, called **coordination conditions on content and causal connection** ("**C5**"). But this was more desiderata than theory, since I did not have enough ma-

[58]For example, to commercial software.

chinery to spell out any additional norms in much intellectual depth. What was needed is what we will address in the second stage of our project: generalisation. That is, we need to consider in detail what sorts of technical generalisations and decisions are required in order to develop this general picture into anything approaching a workable, comprehensive account.

6 Generalisation

No unique set of generalisations is required in order to do justice to participatory systems. Too many consequential subtleties branch out in too many entangled directions to permit accurate cataloguing. Moreover, to do real justice to embodied cognition ultimately requires starting over—building the entire account from the ground up, based on new metaphysics. Still, laying out some of the adjustments and alterations to the traditional conception of representation is a rhetorically and pragmatically instructive exercise. Among other things, it goes some way to illustrate the sorts of issue that a more radical reconstruction will have to face.

In this section, in this spirit, I'll mention a dozen or so such generalisations (see figure 15), grouped into three rough classes:

1. **Participatory:** having to do with the fact that systems "occupy," "inhabit," or are "situated in" their worlds;
2. **Ontological:** having to do with the nature of those worlds that systems inhabit (and the material they are made of); and
3. **Normative:** having to do with the nature of the governing norms to which intentional systems are held accountable.

In a sense, all three depend on a prior, more thorough-going generalisation that permeates everything we have done. I will call it **embodiment**: a recognition that representational systems, and the worlds they inhabit, are constructed from concrete, physical stuff. As we have seen, this fundamental embodiment establishes the powers and limitations of cognitive systems, and undergirds the constituting dialectic between what is and what is not effective. I call it a generalisation of logicism, not just a reconstruction, for reasons that emerged in the last section. Once the foundational conception of traditional logic—especially its bivalent emphasis on formality—is understood concretely, as suggested in the last section, a radically more general picture of representation is unleashed than is traditionally imagined.

But embodiment alone is not enough.

The first "participatory" group of additional generalisations includes several features (besides embodiment) that have been

Generalisations

A · Participatory

1. *Interaction*	· Engagement between system and its environment
2. *Embeddedness*	· Syntactic and semantic domains overlap (limit: fuse)
3. *Context dependence*	· (Weak) Interpretation dependent on context of use
4. *Involvement*	· Orthogonal inside/out & symbol/referent boundaries

B · Ontological

1. *Entanglement*	· Representation and ontology inexorably interrelated
2. *Nonconceptual*	· World doesn't come "pre-parsed" into objects, properties, relations, and other "formal" categories.
3. *Abstraction*	· Commonsense ("natural") ontology—objects, properties, etc.—require abstraction over world's basic, messy, "non-conceptual" structure
4. *Deixis*	· Local, incremental, differential character of physical law implies that content is deictic or indexical
5. *Context dependence*	· (Strong): Meaning dependent on context of use
6. *Features*	· Temporally-indexed features ("it's raining") as a simple form of abstraction (« predicate-object)
7. *Non-discreteness*	· World neither first- nor higher-order discrete

C · Normative

1. *Ends (telos)*	· Generalisation of logic's traditional pair: *Soundness*: Wanting what you get *Completeness*: Getting what you want
2. *Dynamics*	· Interdependence between statical (truth, reference, etc.) and dynamical norms (what to do, how to live).
3. *Objectivity*	· Commitment to the existence of the world

Figure 15 — Generalisations to Logicist Representation

touted as characteristic of "situated cognition." Embracing them will thus give us a handle on many of the traits listed in §■■ as distinctive of an embodied view. The second and third groups, having to do with ontology and normativity, implicate issues that have not received nearly as much explicit attention, at least to date. But these concerns are beginning to make their presence

felt, and (as I hope to show) in some ways they cut deeper into the fabric of an embodied perspective than the merely participatory. Dealing seriously with them adequately requires a more extensive treatment than I can afford here; I will be able to give them just some very introductory remarks.

6a Participation

As we saw, logicism distinguishes two realms: a syntactic realm (*S*), of representational vehicles (such as expressions), and a semantic realm (*D*) or task domain, containing the objects or entities that the representational vehicles are about. Moreover, the governing architectonic took all causal (effective) transitions ('→') to be inferential, understood as an (inferential) relation on the syntactic realm.

INTERACTION

Perhaps the most widely touted characteristic of embodied cognitive systems, taken to distinguish them from logical inference schemes, is the fact that they interact with their environments. So a natural first way to generalise the logicist framework is to license causal connections ('↔') across the *S*-*D* boundary. This move captures an extremely common intuition, that underlies the very notion of perception, and is implicit in such ubiquitous ideas as (i) the standard conception of sensors and effectors; (ii) the "robot reply" to Searle's Chinese room, (iii) the virtual platitude that our senses "connect us to the world," (iv) Harnad's proposal for a generalised "Total Turing Test" to assess intelligence, (v) the imaginative force of a "brain in a vat," thought to be disconnected from any possible semantic realm—and so on and so forth. I will dub it **interaction**: a proposal that effective operations not be limited to system-internal transitions, but include causal coupling across the boundary between systems or agents and the environments they inhabit.

EMBEDDEDNESS

But though it extends pure embodiment, INTERACTION is still too weak. In fact the formulation just given—essentially, of an "inner world" of symbol or thoughts, and an "outer world" that the symbols or thoughts represent, to which it is connected by sensors

and effectors—is a great example of the limitations of a purely amalgamationist approach. For in the act of valorising causal traffic *between realms*, the proposal shares with logic the presupposition that the realms are *distinct*—that the world or task domain that the agent is reasoning is wholly "exterior": outside or beyond the internal realm of mental activity.

The error stems from sundering agent and world. Once the two are conceptually separated, no amount of mere causal coupling is strong enough to glue them back together again.

An example of the difficulties that causal coupling with the environment does not repair arise up in what I have (in another context) called "non-effective tracking": the maintenance, in time, of a dynamic representational state that represents an external on-going process to which an agent is not coupled. This is the sort of thing a creature might do in "mentally tracking" a moving object while it is occluded from visual sight—or that we ourselves often do (badly), after someone has called from the airport and said that they will be home in half an hour, as we imagine them getting into the car, turning onto the freeway, getting to the right exit, etc., so as to be able to predict their arrival. What is striking about such cases is that they involve *non-causal* (i.e., non-effective) *coordination between realms*. In particular, the governing normative conditions on non-effective tracking exploit the fact that the passage of time for an agent, and the passage of time in the agent's task domain, are *one and the same*. They aren't merely "in synch"—in the sense of being two things kept in step by causal coupling. They were never separated in the first place, in any way that would require their being brought back into synchronisation.[59]

Agents are not just *embodied*, in other words, in the sense of being made of concrete physical stuff. They are also *embedded*:

[59]This point must not be confused with the question of the relation between *representing time* and *represented time*. For any dynamic representation of a dynamic phenomenon, those two will be (at least logically) separable. In cases of both effective (standard) and non-effective tracking, the two are as a matter of fact (approximately) coincident: that is what makes them cases of *tracking*. But this issue—a special case of the relation between sign and signified—is orthogonal the relation between agent and world. (See the discussion of involvement, below.)

they live in, are made of, and dwell among the things that constitute their environment. We therefore need a second generalisation, which I will call **embeddedness**: a recognition that the syntactic or effective domain (the stuff of which the system is made, and the agent's "inner life"), and the semantic or task domain (the world the agent represents, the things that it cares about, etc.) will at a minimum overlap, and in the limit be the same.

EMBEDDEDNESS provides for various forms of *coordination* between the realms of representational activity and realms that that representational activity is about. Metaphysically, the point is that not all coordination involves causal or effective coupling.

A striking but familiar example of non-effective coordination is provided by clocks. Clocks are clearly representational: the arrangements of hands on their faces[60] represent what we might call *o'clock properties*: 4:01 p.m., 4:02 p.m., etc.—i.e., properties exemplified by passing metaphysical moments. Clocks were hard to build for exactly the reasons identified in the representational mandate: o'clock properties are indisputably non-effective..[61] It follows that concrete systems can only orient towards them by representing them—by exploiting something else that is effective, that is coordinated with what is not. The task for a clock (or clockmaker) is to exploit the effective properties of the inner workings (clockworks) in order to establish an appropriate relationship between those aspects of the hands that are effectively controllable (the position around the dial) and the non-effective temporal property thereby represented.

The normative conditions on clocks are given in the sidebar on p. ■■ (in brief, clocks are right when the property represented by the position of the hands holds of the metaphysical moment that it is). Needless to say, the temporal conditions on full-scale temporal reasoning and temporal consciousness will be radically more complex. For example, they will involve Husserlian issues regarding the intricate relations between the temporality of perceptual processes and the temporality of dynamic activity thereby per-

[60]I am considering analogue clocks here, though nothing hinges on that simplification.

[61]If "being 4:00" were effective, one could build an automatic kettle that put up water for tea by detecting that the passing moment was 4:00. But of course no such mechanism is possible.

Norms on Clocks

The norm governing the position of the hands is relatively straightforward: at any given moment t, the configuration of hands $_t$ should represent the o'clock property that is true of t. In a sense, a clock has to *track* the passage of time. But it has to track a non-effective property of the passing time; that is what makes the situation representational. In a sense, one can think of the task facing a clock as the dual of traditional inference: whereas inference, at least as traditionally construed, involves moving from one representation of a presumptively stable environment to another, clocks must do the opposite: maintain a stable (at the level of "meaning") relation to a changing environment. Taking the analog (continuous) case as an example, this leads to the following two "correctness conditions" for clocks

(1) correct-speed: $\frac{\partial [\![\odot]\!]}{\partial t} = 1$ (2) correct-time: $[\![\odot]\!](t)$

ceived—intricacies that are necessary in order for systems to authentically perceive the world as on-going and dynamic. The now, the point is only that EMBEDDEDNESS will in general implicate complex forms of coordination and (potentially non-effective) relationality between the effective and the non-effective dimensions of the overlapped (or even unified) "syntactic" and "semantic " realms.

CONTEXT-DEPENDENCE · I

EMBEDDEDNESS opens up the possibility of understanding another of the prominent intuitions underwriting the "situated" movement in cognitive science: a recognition that the representational states of real-world systems are *context-dependent*. Context-dependence is not so much a fact of embodiment per se as it is a semantic consequence of this kind of embeddedness: the fact that material systems are often *located* in their worlds— situated in specific circumstances, in ways that have consequences for their semantic interpretation.

I won't say much about simple context-dependence here, of the sort that characterises indexical expressions (*I*, *you*, *we*, *here*, *now*,

etc.), because it has been so extensively studied. If we make a distinction between a symbol or term's *meaning* and its *interpretation*—where meaning is taken to be approximately the stable, single "rule" or regularity associated with all uses of the term, of the sort that a person acquires when they "learn" the term, and *interpretation* is the context-dependent referent or semantic value that each utterance obtains, on any particular occasion[62]—then this form of context-dependence can be understood as a phenomenon of context-dependent variation for symbols or representations with context-independent meanings. I will name this widely-recognised third participatory generalisation **context-dependent interpretation**. (A more radical form of context-dependence, involving con-text-dependent meaning, will come up in the second group.)

INVOLVEMENT

We still aren't done. Even the conjunction of embodiment, embedded-ness, and context-dependent interpretation does not go deep enough. They potentially (but misleadingly) preserve a sense that cognitive creatures "look out" onto the world—that all the semantic relationships originate in heads, and are directed "agent-external" to the world that we move around and change and dwell in. So in the table I have listed a fourth and final participatory generalisation, labelled **involvement**.

The aim of INVOLVEMENT is to recognise that semantic directedness ('⇒') and causal coupling ('↔') are *orthogonal*.

To understand what this comes to, note that representational (computational, logical) systems can be understood in terms of two distinctions or "boundaries":[63]

[62]So when you use the term 'I,' the interpretation is *you*; when I use 'I,' the interpretation is *me*. Thus our interpretations differ. If each of us meet someone we have never met, and they use 'I,' the interpretation is *them*—a new interpretation, one we have never before encountered. But we are not mystified, when hearing that new person say the word, because we know *what the word means*. (Thus meanings can rather glibly be viewed as a rule or regularity of the form λcontext.interpretation.)

Put it this way: dictionaries give meanings, not interpretations. That is why there is only one entry under the word 'I'; not ten billion, one for each person in the universe.

[63]The discussion is this section is a radically brief summary of some of the

1. A **semantic** boundary, between a representational vehicle and its referent (what it is normatively oriented towards); and
2. A **physical** boundary, between a system's insides and outsides.

Given these boundaries, one can then identify a pair of theses on which the classical model is based:

1. An **alignment** thesis, claiming the 2 boundaries line up; and
2. A thesis of **isolation**, claiming that the 2 (allegedly-aligned) boundaries are something of a moat (causal, logical, explanatory).

Jointly, these two theses entail that all of the symbols or representations lie within the system, and all of the referents are to be found on the outside (roughly what was suggested in the original logicist figure 3). What the INTERACTION generalisation does (the idea underlying the robot reply to Searle, the idea of extending an inferential model of cognition by adding sensors and effectors, etc.) is to deny **isolation**: the idea that transactions the boundary between the symbol system and the "outside world" is closed. As far as it goes, as we have seen, that is surely correct, for any plausible notion of an embodied cognitive creature. But what that analysis fails to recognise is that **alignment** is false as well: *the boundary between symbols and their referents, and the boundary between the inside and outside of a system, are orthogonal.*[64] Not only are there (in the real world) internal symbols with external referents, as imagined on the classical image (thoughts about a friend or enemy), but also internal symbols with internal referents (introspection and self-knowledge), external symbols with internal referents (the advice of friends and psychiatrists), and external symbols with external referents (roads signs directing you to the airport). Plus, there are causal transitions *between and among all of*

---

results of AOS-II.

[64]This is one of the primary results of the analysis of formal symbol manipulation «ref AOS-II».

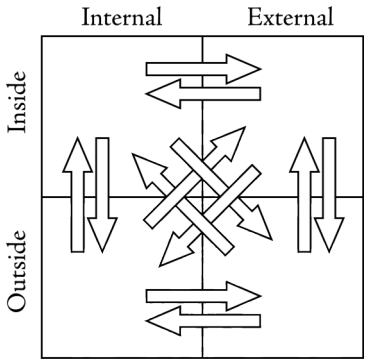

Figure 16 — Participatory Transactions

these four kinds. Figure 16 gives an indication of the structure of this terrain, with four kinds of representational example, and sixteen different types of thereby engendered causal transition. For example, a plausible normative constraint on the process of *reading* might be not that one *represents* the text being read (as many classical analyses suggest), but rather that one "internalise" the text, by constructing internal representations whose semantic content is the same as that of the external representations with which one is interacting (see figure 17).

I make no claim that even these four mandates are enough to ensure the kind of "being in the world" that everyone takes to be constitutive of a situated, embodied view. But as we will see, they are enough to cause profound consequences to the theoretical frameworks in terms of which to understand the overarching representational mandate, of local effective processes governed by overarching but non-effective norms.

6b Ontology

One feature of the separation of realms (as we have seen) is characteristic of the logicist picture is its foundational ontological presupposition that the character of the semantic realm (what objects, properties, relations, etc., constitute it) and the character of the syntactic or effective realm are established independently—and also "extra-theoretically," in the sense of being assumed to be fixed, prior to and independent of the characterisation of the agent as cognitive. This structural character reflects, in technical guise, a guiding simplification that undergirds much of the analytic philosophy on which traditional cognitive science rests: an assumption that the theory of representation

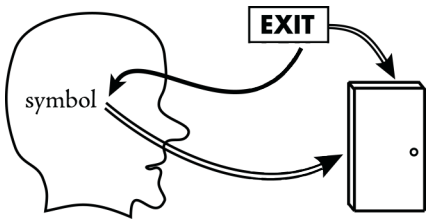

Figure 17 — Norms on Reading

(how creatures take the world) and the theory of ontology (what the world is like) are *independent*.

If any general theme underlies the sorts of shift I am recommending as necessary in order to do justice to an embodied perspective (beyond the explicit emphasis on concrete materiality, participation, etc.), it is a move to dismantle and defuse all sorts of sharp independence underwriting traditional logicism.[65] No independence goes deeper than the just-alluded-to separation of representation and ontology. More strongly, though nothing we have said so far argues for it, I want to start with a move to generalise—to be honest, to deny—this metaphysical assumption. That is, I want to endorse what I will call **entanglement**: a recognition that representation and ontology interrelated—that how we represent the world to be, and what the world is like that we thereby represent, cannot be given independent explanations.

Three immediate comments.

First, it is critical to realise that embracing ENTANGLEMENT is (in and of itself) no endorsement of radical solipsism, idealism, or any other metaphysical stance that *fuses* representation and represented. As mentioned earlier, the idea that if two things are not the same, then they must be independent, is *exactly the kind of ideological commitment to independence and sharp distinctions that I am at pains to deny*. All that is claimed, by the ENTANGLEMENT mandate, is that a fully general (rehabilitated) approach to representation must allow representation and ontology to be at least partially interdependent.

Second, in spite of those very mildly conservative observations (essentially, recognising that some vaguely realist intuitions must or at least may need to be retained), I would be the first to admit that dismantling the analytic assumption that ontology and representation are independent is an *extraordinarily expensive move*. The theoretical consequences are staggering, with implications that shake the very foundations of what it is to do science, to give a theoretical account, to know. Just a few of the most evident consequences will be touched on below; but it is in recognition of

[65] In AOS I take this pervasive sense of independence and sharp distinctions to be the deep meaning of formality (one that reaches much further into the conceptual bedrock of logic, science, etc. than the more superficial positive and negative readings we talked about in §■■).

sequences will be touched on below; but it is in recognition of the full power of ENTANGLEMENT's implications that I said above that the only ultimately palatable way to do justice to embodied or situated cognition will require complete metaphysical overhaul.

Third and finally, in spite of the expense, it is impressive how many different currents and voices in contemporary cognitive science argue, implicitly or explicitly, for exactly such a loosening of the traditional assumption, and a potential melding or meshing of representational and ontological concerns.[66] What these arguments and these voices together imply, I believe, is that embracing ENTANGLEMENT—and recognising that cognitive science must take on blatantly metaphysical and ontological issues—is the most urgent intellectual issue that faces current cognitive science and philosophy of mind.

ENTANGLEMENT is such a strong mandate that it is perhaps almost fatuous to list any others. But in the table I have enumerate five more, to give a flavour of the sorts of ontological task that await us:

1. **Entanglement:** allowance for the fact that representation and ontology inextricably interrelated
2. **Nonconceptuality:** a recognition that the world does not come "pre-parsed" into the theoretically-familiar categories of objects, properties, relations, sets, states of affairs, propositions, possible worlds, and the like, as assumed in the logicist tradition.

   I label this NONCONCEPTUAL because it is a theme of the literature on nonconceptual content[67] that *conceptual* content is content that takes the world to be structured in this way (i.e., in terms of objects, properties, relations, etc.), opening up the possibility that nonconceptual content might be content that takes it in some other way. I have argued elsewhere[68] that the warrant for nonconceptual content is ultimately ontological, not epistemological—i.e., that the raison-d'être of nonconceptual

[66]«Cite Thompson, Varela, & Rosch; Lakoff and Johnson; Haraway; Lave; Chemero; Cussins and the non-conceptual literature; Objects, etc. »
[67]«Ref Evans, Cussins, Bermudez, Peacocke, etc.»
[68]«Ref "The Nonconceptual World"»

content lies ultimately in the *world*. It is reality that is not aboriginally conceptual, that is (i.e., that is not structured in the way in which conceptual content takes it to be).

Many will assume that this ontological version of nonconceptuality is a species of *non-realism*. But that is a legitimate label only on an assumption that the conception of the world as conceptually structured (i.e., as consisting of objects, properties, relations, etc., as classically imagined has some incredible kind of pre-metaphysical claim of priority. If one assumes, as I do, that the world is not autonomously so structured, then it is *conceptual* content that flirts with be irrealist, not nonconceptual content.

3. **Abstraction:** a recognition that the commonsense ontology represented by conceptual content, as described above, involves profound capacities for *abstraction* in cognitive creatures, which cognitive science needs to explain.

4. **Deixis:** a recognition that fundamental facts about the nature of physical existence, having to do with the incremental, differential character of physical law (i.e., the ontological facts that warrant our expressing the laws of physics in the form of differential equations) imply, as a consequence, that any physically possible form of representation will (at least in any simple form) be originally *deictic* or indexical.

   It is not first-person content that is mysterious, from a physicalist point of view, in other words. Rather, the mystery—the theoretical puzzle that challenges cognitive science—is how concrete agents can achieve third-person reference or content.[69]

5. **Strong context-dependence:** a recognition of the possibility that not just interpretation (in the sense described in §■■), but *meaning as well*, may be context-dependent. It is not just that different utterances of 'I', 'now', etc., have different interpretations on different occasions, in ways governed by a stable, context-independent regularity. Rather—at least in general—it may be that *even what ordinary words (or cognitive symbols) mean* may depend on

[69]«Ref "Who's on Third?"»

contingent or circumstantial facts about the situation in which they are used.

I call this form of context-dependence *strong* because, as with the other ontological generalisations being listed here, its implications for metaphysics are strong. But there is nothing intrinsically contradictory to the idea—or, necessarily, irrealist. Some may assume that if meanings are not entirely fixed, then they must be completely fluid—thereby taking leave of any possible realist commitment (and in the process vitiating any talk of world-directed norms). But to think that is merely another instance of a black-and-white assumption. There is no logical reason why the meaning of words cannot (in general) be partially fixed, or at least relatively stable, but nevertheless be partially bent and shaped as well, by contingencies of a discourse situation.

6. **Features:** a recognition, in consort with above-mentioned suggestions that commonsense ontologies of objects, properties, etc., may involve sophisticated conceptual abilities, that a simpler way of registering the world, in terms of (temporally-indexed versions of) what Strawson has called *features*, more like property- or relation-instances than anything with the full logical structure of objects and properties, may figure in nonconceptual representational schemes.
7. **Non-discreteness:** a recognition that the structure of the world may not in general be *digital* or *discrete*, either in the ordinary sense (in which people think computers are digital), or in the sense that Haugeland has called "higher-order discreteness"[70]—a kind of clean separation of *concepts* and *properties* that is familiar in mathematics and science, but seems radically unlikely to hold of such everyday notions as (for example) confidence, ego, chutzpah, bravado, arrogance, braggadocio, etc.

Needless to say, these are mere telegraphic labels of subjects that would require vastly more space even to convey an adequate sketch of. But they illustrate the sorts of ontological reconfigura-

[70]«Ref "Analog and Analog"»

tions of the world that we, as cognitive scientists, are going to have to deal with, if we take the embodied, participatory stance seriously.

6c Normativity

We have said little, so far, about normativity. But as mentioned in the discussion of logic, to enter the realm of representation—description, language, interpretation, truth, etc.—is to enter a world of phenomena governed by asymmetric (paired) evaluative predicates: true vs. false, good vs. bad, working vs. broken, beautiful vs. ugly—where in each case one option is *better*, or *more worthy*, than the other. Accurate descriptions are better than inaccurate ones; information is better than misinformation, helpful behaviour is better than unhelpful behaviour—and so on. In fact one very plausible definition of intentional systems is that they are just those systems that are subject to norms.

The question is what to say about how to generalise the normative structure implicit in logicism in such a way as to incorporate the full range of norms that are appropriate to embodied cognitive agents.

For starters, we can generalise soundness and completeness in terms of a more general characterisation of **ends**. As described above, once states of a system can both engender causal consequences, in virtue of their effective structure, and stand in some kind of semantic relation to (potentially distal) states of affairs, the issue arises of whether, if an operation happens, or behaviour takes place, the result does or does not meet any applicable governing norm. There are, in general, two ways to fail. This leads to a natural reconstruction of the two traditional norms on inference:[71]

1. **Soundness:** wanting what you get
2. **Completeness:** getting what you want

The more substantive question has to do with what it is you want.

To get at this, consider logic's emphasis on truth. Truth is justifiably famous—but not particularly general. Within the logicist

[71]This informal but perspicuous formulation is due to John Etchemendy.

framework, moreover, it has been treated as a **static norm**, in the sense of applying to (passive) *sentences* or *claims*—i.e., to *states*.[72] Full-blooded intentional systems, however, are *dynamic;* hence governed by **dynamic norms**—norms that govern process.[73] In logic, the operative dynamic norm is derivative—defined in terms of a static norm. Reasoning, deduction, inference to the best explanation, etc., are all mandated to *preserve* or *produce* truth or explanation, where it is (critically) assumed that what it is to be true, and what it is to be an explanation, can be defined independently of, and prior to, the processes of their preservation or production.

This explanatory strategy—of starting with a (presumptively autonomous) static norm, and then defining dynamic norms in terms of it—has been picked up by other intentional sciences. Economic models of rationality and decision-making, for example, often use the dynamic norm of *utility maximisation*—where utility is (once again) presumed to be static, prior, and autonomous. But the general strategy of defining dynamic norms in terms of static norms doesn't generalize. And no computer scientist believes it. On the contrary, what practical experience with computing has taught us is that you it is vastly more general to proceed in the opposite direction: taking the semantic content (meaning) of a symbol or expression or data structure to be determined (even to exist) depending on *how it is used*—i.e., on the role it plays in the overall system of which it is a part. Rather than *define dynamic norms in terms of static ones*, that is, programmers *define static norms in terms of dynamics ones*—in a (perhaps unwitting) endorsement of the Wittgensteinian maxim that "meaning is use." And so this I have listed as our second normative generalisation: that we shift our original explanatory dependence from static to dynamic norms.

If we get our static norms derivatively from our dynamic ones, where do we get the original dynamic norms? What are they like?

[72]By static norms I don't mean norms that don't change, over time; evaluative metrics on book design, or on human beauty, may evolve considerably, but would still be counted as static, on my typology, because what they are evaluative predicates on—books or motionless bodies) are essentially static things.

[73]«Say: should be (or change to): 'statical' and 'dynamical'»

What governs, what puts value on, what evaluates, the use—i.e., the life and times, the activity—of general intentional processes? Though the question isn't usually asked so baldly, a variety of alternatives are being explored in contemporary cognitive science. But one dynamic norm is currently receiving by far the most scientific attention—in cognitive science, ALife, evolutionary epistemology, research on autonomous agents, and biology: **survival**.

It is clear how to get a norm out of survival: a process or activity is deemed *good* to the extent that it is *adaptive*—i.e., to the extent that it aids, or leads to, the long-term survival of the creatures that embody or perform it. This idea of resting normativity on evolution is seductive. It has been used to define a notion of *proper function*, for example, in terms of which to decide whether a system is *working properly* or is *broken*. Thus the *function* of the heart is to pump blood, and not to make a "lub-dub" sound, because hearts were evolutionarily selected for their capacity to pump blood, not for their sound-making capabilities. Similarly, the function of sperm is to fertilize eggs because that is why sperm have survived (even if only a tiny fraction of them ever serve this function).

Most interesting for our purposes, however, is the use of this same idea to define semantic content (meaning, reference, representation, truth). The representation in the frog's eye *means* that a fly is passing by, some people claim, because it leads the frog to behave in an adaptive way towards that fly (namely: to stick its tongue out and eat it) in a way that contributes to the frog's (not the fly's) evolutionary success. Similarly, the shadow on the ground conveys information *about* the hawk in the sky to a mouse just in case it plays an evolutionary adaptive role of counterfactually covarying with the presence of hawks in a way that allows mice to escape. That is, modern philosophy of mind has begun to change from logic, in taking the static norm of reference and truth to derive from the dynamic norm of leading to an adaptive or evolutionarily successful life.

Have we reached the end of the line? Will evolutionary survival be a strong enough dynamic norm to explain all the norms that apply to cognitive agents: justice, altruism, authenticity, caring, freedom, and the like? Personally, I doubt it. But in a way that is just the point. For what is at stake, for cognitive science, is

not what will ultimately sub-serve the norms we need in order to understand human activity, but to understand *what the dynamic norms are in terms of which human activity is conducted and understood*. And that, I hope, is obvious: dynamic norms on human activity govern *what it is to live*—what it is to live well, to do good, to be right. That is: **ethics**. And not just ethics, but whatever governs whatever you do: ethics, curiosity, eroticism, the pursuit of knowledge for its own sake…and so on and so forth, without limit.

In sum, taking on full-fledged dynamic normativity is an unimaginably consequential move. It implies that any fully rehabilitated account of representation—any transformation broad enough to incorporate arbitrary embodied and embedded intentional systems, and thus to treat meaning along with matter and mechanism—will also, thereby, **have to address mattering as well**. Put it this way: in spite of logical practice, it won't generalise to bite off *truth* and *reference*, and glue them, piecemeal, onto physical reality, without eventually taking on the full range of other norms: *ethics*, *worth*, *virtue*, *value*, *beauty*. By analogy, think of how computer science once thought it could borrow *time* from the physical world, without having to take on *space* and *energy*. It worked for a while, but soon people realised what should anyway have been predictable: that time is not ultimately an isolable fragment—not an "independent export"—of physics. By the same token, it would be myopic to believe that the study of intentional systems can be restricted to some "safe" subset of the full ethical and aesthetic dimension of the human condition—and especially myopic to believe that it can traffic solely in terms of such static notions as truth and reference, or limit itself to a hobbled set of dynamic norms (such as survival). To believe that would be to be an ostrich, not a hero.

Moreover, to up the ante (in case this all seems too mild), something else, if anything even more expensive, is implied by these same developments. (Moreover, this is where the story starts to fit together, though it is also what mandates the development of new metaphysical foundations.) I said above that the classical model assumed that the meaning of symbols and representations could be assessed in terms of the objects and properties in the world that they corresponded to, independent of how those

symbols and representations were used. But I also said, in the discussion of ontology, that many modern cognitive scientists no longer believe the classical model—in part because the physical world does not supply the requisite objects. That means, as we have already admitted, that it is incumbent on a theory of representation to explain the objects that figure in the (conceptual) content of a creature's representational states. What we didn't say in that ontological discussion, however, is that those objects are to be explained in terms of the normative structure governing the representations whose contents contain them. And those norms, we have just admitted, are ultimately grounded on *dynamic activity*.

It follows that the *material ontology of the world*—what objects and properties there are, for a given creature (not just what objects and properties the creature *takes* there to be, but what objects and properties there *actually are*, in the world, for that creature)—will, on the generalised account, be seen to be a function of that creature's projects and practices. For high-level social entities this isn't surprising: date-rape didn't exist, I take it, for the aboriginal singers of the Australian song-lines; the strike zone (a favourite object) isn't part of the furniture of the world, for earwigs. But the present claim is more radical: it suggests that what is true for date-rape and strike zones is true for food, clothing, rivers—perhaps (who knows?) even for the number four.

Ontology is inextricably linked to epistemology, in other words; that much we said with ENTANGLEMENT, above. What we are adding, now, is that epistemology is inextricably linked to ethics. These are conclusions I am happy with; but they are nothing if not strong. What is striking about them in the present context is that we have come to them by making two seemingly innocent moves: (i) by understanding that material ontology involves conceptual abstraction; and (ii) by giving dynamic norms explanatory priority over static ones.

We can summarise this conclusion etymologically.

A material object is something that matters

It must matter, in order for the normative commitment to be in place for the objectifying creature to take it as an object: to be committed to it as a denizen of the world, to hold it responsible

for being stable, obeying natural laws, and so forth—and to box it on the ears, when it gets unruly. It is no pun, in other words, or historical accident, that we use the term 'material' as a term for things that are concrete (made of "matter") and also as a term for things *that are important*—as in 'material argument,' or 'material concern.' In fact that is one way to see where the embodied cognition movement is headed: whether it knows it or not, it is going to have to heal the temporary rift that for 300 years has torn matter and mattering apart.

7 Application to embodied cognition

One task remains. We need to understand how our proposed rehabilitated model deal with embodied cognition. That, after all, was our original goal: to combine the best in representational and nonrepresentational accounts, in order to avoid the fundamentalist excesses of figure 1.

Three preliminary remarks.

First, it is important to be clear on the question being asked. Many discussions of the relation between "new" and "old" cognitive science compare a proposal for a new "embodied" approach to representation as traditionally conceived. Thus van Gelder and Port contrast a dynamical systems approach, which they recommend, to their conception of the classical model (which they call "computational"), which they criticise. That is not the contrast I am addressing.[74] Rather, setting aside any vestige of the classical view, now, I want to understand the relationship between:

1. Various proposed *non-representational alternatives*, of which dynamical systems theory is one candidate, though there are others; and
2. The rehabilitated conception of representation being developed here.

For what kinds of system is each framework most appropriate? What kinds of insights and understandings are expressible in each framework's terms? What kinds of behavior warrant the admittedly more complex analysis provided in terms of a reconstructed notion of representation? How well does the rehabilitated notion of representation deal with the ■■ characteristics cited in §■■ as distinctive of the embodied view? Those are the sorts of things we want to know.

Second, it's not 100% clear what a "dynamical system" is. At the most general level, dynamical systems theory is a body of mathematics, applicable to any situation in which a system which can be described in terms of temporally-varying instantiations of

[74] I am especially not interested here in the issue of whether their view can legitimately be called computational—which I think it cannot.

measure properties—causal, semantic, emotional, whatever. By itself, that is, nothing in the term "dynamical system" necessitates the characterised properties being in any sense physical or effective. Thus a committed Cartesian could talk about God's waning love in dynamical terms (figure 18). I take it that the presumption in cognitive science, however, is that a dynamical systems account of a system's behaviour is understood to be a description of its causal (effective) behavior. That is what I will assume in the following.

$$\frac{\partial(\text{love-of-God})}{\partial t} < 0$$

Figure 18 — Dynamics

Third, a reminder about the (non-effective) nature of semantics. As we have said, it is a something of a metaphysical theorem—at least for physicalists—that systems work, mechanically, solely in virtue of their total effective (causal) structure: the effective structure of their internal arrangements, in interaction with the effective (causal) structure of the environment they are embedded in. This is a general claim, which holds of absolutely everything that there is: representational and nonrepresentational alike. So the following is not a possible objection to a representational (or other kind of intentional) analysis: "What do you mean, semantics? All that exists, for this system—all that there is—is a pattern of causal transitions and structural couplings to the immediate environment! How could there be anything else? Look at the system; attach any instruments you can devise. Show me something more than that!" This objection fails because it clearly assumes (e.g., in its reliance on *instruments*) that "all that exists" means "all that exists, *causally*"—all a meter could detect, all that involves the expenditure of energy, all that traditional sciences recognize as real. But all parties agrees with that; that was the exact import of our reconstruction of the negative reading of formality as a claim that semantics is not effective. We have already admitted that semantics cannot be detected by a (causal) instrument. To suppose that it *could* be would be to suggest that representation violates physicalism, which no one is suggesting.

Rather, what the representationalist (intentionalist) is claiming is something else: that an account of a system's local, causal interactions *does not exhaust the constitutive facts about that system*—the facts that would need to be accounted for by an

facts that would need to be accounted for by an explanatory theory.[75] For remember what we said about semantics: they operate as *non-effective governing norms*. In order to show that a system is not semantical, therefore, one must show that it is not so normatively governed. That is not quite as easy to do as a simple causalist might imagine.

7a First pass • Formal

The first thing to say is that the reconstructed representational account we are sketching is extraordinarily broad. Indeed, all it really comes to, so far, is that a local, causal, effective account must be given, of how the system works; plus a potentially non-local, non-causal, non-effective account of semantic interpretation; and that the two be tied together by constituting norms. By hypothesis, the view of dynamical systems we have endorsed is merely one way of giving the first of these: a causal account of behaviour.

Being a dynamically-described causal system, however, by itself has *no bearing whatsoever on whether the thereby-described system is representational*. That is because, from the point of view of pure mechanism, the new representationalism imposes no apparent constraints! Representation, as we said at the beginning, is (in its

[75]By analogy, think about all the possible cursor shapes that can be displayed on your computer. On most operating systems, cursors are arbitrary 16 × 16 bit binary patterns, which a program can set arbitrarily, so as to draw the familiar shapes we all know: arrows, hourglasses, cross-hairs, etc. Since there are 16^2 or 256 bits, each of which can be on or off, there are $2^{256} \approx 10^{77}$ different possible shapes—or about 100,000 times as many as there are electrons in the universe. Of these, we use a few hundred, or at most a couple of thousand.

Suppose one wants to provide a theory of cursors. One theory might simply say that cursors are 16 × 16 bit patterns, and describe how they are set and manipulated by hardware and software. In terms of the local pattern of causal behaviour, that account may be complete. But something may be left out. For example, suppose (falsely) that the only cursors that are ever drawn are shapes that resemble naturally-occurring artifacts. A full theory of cursors, therefore, would have to include a theory of *what it is to resemble a naturally-occurring artifact*. That additional theory would not be a theory that added or changed—especially "in the small"—how the cursor works, causally. But it would nevertheless reconstruct constitutive patterns of cursors that the purely causal story would not.

full potentiality) an extraordinarily broad notion. So the question on the table is going to boil down to the following: in what circumstances is it productive—valuable, explanatory, and *true*—not only to give an account of how a system works, mechanically, but to tie that (normatively) together with an account that interprets the system?

More strongly, the conception of representation that we have been developing was explicitly designed to include the ability to treat the sorts of behaviour that embodied cognition takes to be essential. In particular, consider the list of eight contrasting pairs of properties, listed at the beginning of §2g (page 10), of what distinguished embodied cognitive systems from classical (allegedly "computational") ones. Of these, the first—a shift from pure abstraction to concreteness, or an endorsement of the importance of EMBODIMENT—has not only been dealt with, but has underwritten the entire story we have been telling—about effectiveness, representation's raison-d'être, etc. Whatever else is true of our reconstruction, in sum, it puts concrete materiality squarely on center stage.

The third (I will return to the second in a moment), that the system NOT BE SEPARATED from its semantic realm—is part and parcel of what we dubbed a participatory view (cf. earlier remarks on perception, action, tracking, introspection, cross-cutting boundaries, etc.). Similarly for the fourth requirement, that a system be dealt with as ENGAGED with its environment. Finally, the fifth and sixth requirements—that we deal with DYNAMICS, and treat CONTINUOUS behaviour—have also been made room for (both were illustrated, for example, in the discussion of clocks, including the "clock" equation). Similarly, the sixth characteristic, that embodied systems be understood as CONTEXT-DEPENDENT, has been fully embraced. Not, let me hasten to say, that the postulated reconstructive framework provides theoretical tools for dealing with any of these aspects. On the contrary, tremendous work remains to be done to understand how to treat such features adequately. The point is only that there is nothing in a representational approach, per se, that stands at odds with any of them.

The eight listed characteristic was its selection of NAVIGATION, rather than deliberative, ratiocinative thought, as the paradig-

matic "cognitive" activity of an embodied view. As I hope is clear, this is not a *requirement* of the reconstructed view; its aim was to be neutral on such decisions—providing the wherewithal to treat of both thought and navigation (and a host of other activities). So while the requirement is not exactly met, nevertheless I count this greater generality a feature. And in a sense the same is true of the previous three: while continuous, dynamic, context-dependent representations have been embraced, nothing prevents the treatment of discrete, static, or context-independent ones. This counts to the view's benefits: its aim was to be catholic, able to deal with the full range of possibilities, not to take an ideological stand on either side.

Turn then back to the second pair, having to do with the linguistic, explicit nature of the representational vehicles, on the classical side, which were rejected on the dynamical side. This is a somewhat subtler case. There are two distinct issues at stake.

The first has to do with how the reconstructive account deals with content. As indicated in the discussion of ontological generalisations, it is no part of representation, as we have reconstructed it, to be especially committed to explicit, conceptual, or linguistic content. On the contrary, we have made explicit gestures towards non-conceptual content, which stands as a strong candidate for a form of non-linguistic or non-explicit content. But as in the previous cases, the aim for the representational *framework* is for it to be neutral on the question—exactly so as to allow the theorist to explore different kinds.

The second issue does not have to do with content that is not linguistic, but rather with systems or behaviours that *do not have content at all*. That is, how are we to treat systems that (in spite of the breadth of our rehabilitation) are *genuinely non-representational*? It can hardly be counted against the rehabilitated account that (by itself) it does not deal with them; that was not its aim. For cognitive science, though, we do need to understand the powers and limitations of non-representational systems—which finally brings us back, full circle, to the first strategy mentioned at the very outset: of amalgamation.

7b Second pass • Substantive

Formally, we have concluded, nothing in the list of characteristics

of embodiment militates against a representational account. But that is an admittedly thin result. After all, the rehabilitated account was expressly designed to accommodate this list. The remaining question is the substantive question: what (given that we have gone to all this work) does a representational account buy you—and when are such analyses warranted?

This, finally, is where the rubber meets the road.

Start with the most basic stipulation of the "embodied cognition" movement: that cognition has evolved in response to, and must be understood in terms of, the material conditions and capacities of the cognizing organism. *Start with the body.* The body is a mechanism. So am embodied approach must start with the mechanical—which is to say, effective—capacities of the organism. This much is gospel.

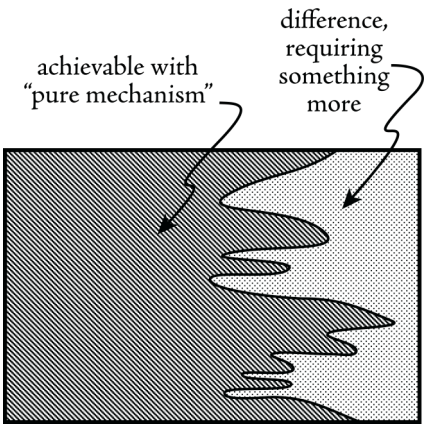

Figure 19 — Brooks' Challenge

By a **purely mechanical system** I will mean a systems whose constitutive regularities are exhausted in terms of the causal/ effective interior structure, and the causal/effective relations that it bears to its environment. Physics, and its immediate higher-level natural sciences, such as chemistry, thermodynamics, etc., I take it, study purely mechanical systems.[76] Dynamical systems theory, as we are characterising it, developed as a mathematical framework in terms of which to analyse the behavior of such systems. As Bechtel has noted, dynamical systems equations are in a sense covering law equations, more than mechanical accounts of how the systems work—a fact that will prove to be of some importance, in a moment—but for now we can continue to assume that the regularities that the dynamical equations account for are behavioral regularities, regularities that *have* an (immediate) causal explanation.

One of the insights of the embodied cognition movement—

[76]Literally: they study phenomena as purely mechanical.

reaching back as far as Raibert's pogo-stick robots[77]—is that we do need to understand bodies, and their natural dynamics, mechanically. As much was admitted on even the simple amalgamationist research strategy with which we started. Something else researchers have repeatedly discovered—epitomized in Braitenberg's book— is that as astonishing amount of behavior can be generated merely by placing a mechanism, of some functional or causal capacity, into a structured environment.

Moreover, it is not just that a great deal of behavior can be so explained, but for reasons of economy, evolutionary plausibility, and sheer good sense, it is best to try to explain as much behavior as one can, in this way. This strategy has been explicitly endorsed by Rod Brooks, who formulates it as something of a maxim:

> *Explain everything you can purely mechanically. Only use representation for the "residue"—for that last increment of cognition that cannot be explained purely mechanically.*[78]

This strategy is figuratively depicted in figure 19. The overall rectangle is meant to indicate the full suite of capacities required for general intelligence; the white central region indicates the range of capabilities that can be explained in purely mechanical terms.[79] The shaded re-gion—the difference between the two—is meant to indicate Brooks' "delta" or "residue"—the range of capacities that do require, for their deployment, representational powers.

The way we can get at our question, therefore—of what it is that generalised, reconstructed representation is good for—is to inquire about the nature of the white region, and the nature of the shaded "delta." That is, we face two questions:

1. What can be done with a pure mechanism?
2. What requires the additional resources of representation?

And finally, we are ready to reply.

The answer was implicit in §■■'s discussion of what is and what is not effective. Remember that the constraints of materiality or mechanism are the constraints of *physical being*. More particularly, they are the constraints of *effectiveness*—that was the

[77]«Ref»
[78]«Is there a quote I can use? Check his article in Mind Design II.»
[79]I am not making any supposition in the area of this inner curve.

ticularly, they are the constraints of *effectiveness*—that was the whole point of identifying effectiveness as a critical subject matter. But what is effectiveness like? And what can it do? Well, among other things, as we saw, effective properties are local properties, due to the fundamental locality of physical law. That leads to the following general claim. What can be done, purely mechanically—and what can be explained, therefore, purely mechanistically—are two things:

1. Regularities having to with effective properties of the system itself (i.e., its inner constitution), and
2. Effective properties of the environment in which the system is deployed.

But what are the effective properties of the environment? They, too, are intrinsically local. It follows from the nature of physical law, that is, that:

> *With respect to pure effectiveness, what a system can deal with, mechanically, is its own (internal) effective state, and whatever impinges on its surface.*

The picture, in other words—and by no means is this surprising—is very much along the lines of that of Maturana and Varela's **structural coupling**. A system (according to them) consists of an organised amalgamation of parts, whose effective properties come together to give the system some behavioral repertoire, which is then "coupled" into the immediate environment. What the system "does," as a result, is: (i) potentially adjust its effective internal arrangements (i.e., adjust its "state"), and (ii) potentially adjust (push and pull) on the impinging lamina of forces and fields that press in on its surface. Except that the "pushing" and "pulling" are symmetrical: neither affects the other any more than it affects them. This is why the Maturana/Varela image is apt: a system adjusts its internal state, and is "structurally coupled" to its environment. Re pure causality or pure effective mechanism, *that is all*. That is all that is going on. And given our background assumptions of physicalism, so long as we focus only on causal aspects of the system, that is all there is to *any* system. The locality of physics prohibits more.

We can summarise this as something of a maxim:

> *The life and times of a purely mechanical system is wholly and entirely exhausted by what happens to its internal effective arrangements, and what happens at its immediate periphery.*

That's all.

Two points.

First, not only purely mechanical systems, but all systems *qua* mechanical systems, *are always 100% coupled to their environments* (in this sense). The nature of the environment may change. But whether the system is coupled to it may not. The reason is simple: physics does not allow disengagement.

Second—and this is what matters most—it is not the *world* that such systems are engaged with. Rather, what they are coupled to a 3dimensional laminar surface of forces and fields, pokes and pressures, that is literally and constantly in the system's face. Qua physical mechanism, that is, there is no door over there across the room, no room downstairs, under the floor, no food around the corner in the cafeteria, no warm and snuggly bed, back home. *Those things are distal.* And distal things are inaccessible, as such, to pure mechanism.

So we have the answer to Brooks' paired questions.

Start with the first. What can you do, purely mechanically? The literal answer is this: you can deal with what is purely effective. What is purely effective is constrained, among other things, to be what is entirely local. So what you can do, purely mechanically, is (at most) deal with what is purely local—locally onboard you, or locally right there at your periphery. You can't even deal with *everything* that is local. Only with that vanishingly small percentage, overall, of those local properties that happen to be effective.

7c The role of representation

Is dealing with what is local, and effective, the sum total of intelligence? No. Part of what it is to be a cognizing creature is to inhabit, live in, deal appropriately with, *the world*. Perhaps not the entire world, to start with—maybe just a bit of the world, around

your natural habitat.

What representation is *for*, therefore, is to *deal with the world*. To know that there is a universe out there! To deal with what is distal—with things not at your immediate proximal periphery, but some distance away: across the room, down the street, around the corner. To understand that things don't cease to exist, outside the door, around the corner…at the limits of your senses.[80]

Look around you. What do you see? It's amazing—you see chairs, tables, people, perhaps; maybe a mountain or a stream. Perhaps the inside of a car. None of these things is at your periphery. In fact—stunningly—you *can't* see anything, if it is pressed right up against your eyeball. That is because the content of our experience is the world *is at the end of those double-tailed arrows*.

Not only is experience representational, in other words, but the content of experience is invariably something we are *not* coupled to.

The final answer of this long journey, that is, is something of an ironic opposite to that proposed by Maturana and Varela. Qua pure mechanism, they are right: what it is to be a mechanism is to be structurally coupled to a manifold surround. But that is not what the world is *like* for a cognizing creature. Re what it is *like*, logicism was closer to the answer. What you represent—what you think about—is *not* what you are coupled to ('↔'), not what is effective, but what you are semantically and normatively oriented towards ('⇒')

What the world is *like*, that is—for us, and for any system that represents—is *how we represent it as being*, where to represent is to exploit the plasticity of that same causal coupling and locally impinging surround, so as, without violating the pre- and proscriptions of physics (it really is a magic trick) to stand in appropriate relation to what one is *not* causally coupled to. Moreover, it is exactly that fact that we are oriented towards the world, conscious of the world, committed to the world, that makes us intelligent.

[80]As Strawson put it: "How do we know that our senses fail, rather than that the world fades?" '«Ref»