Introduction

The significance of computational thinking has arguably been as great as the impact of the technology itself. The influence is apparent in many fields, including physics, chemistry, biology, psychology, linguistics, economics, and sociology. Computation has even influenced mathematics, where theories of computational complexity and possibilities of superhuman calculation have had an impact on the very idea of proof. And disciplinary influences alone do not convey the extent to which computational ideas have permeated our overall theoretical attitude. They now permeate our general metaphysical outlook—on ourselves, the universe, and our place within it.

Nowhere is computation's influence more noticeable than in the adage that we have entered "the age of information." Access to vast online information resources are computationally mediated. The information processing model of mind, which construes intelligence as computational processing of information structures, has revolutionized psychology and underlies artificial intelligence. Computational metaphors of sensory bandwidth, processing power, and memory capacity are so ubiquitous that they are barely theoretically visible. Communication is widely understood as the conveyance of information, learning as its acquisition, and knowledge and thinking as its deployment. Though the nature of information is contested, as witnessed by divergent theories, accounts of information processing ultimately rest on forms of computational thinking.[1]

[1]There are two main thrusts in attempts to define theories of information. Most familiar to computationalists are quantitative accounts that define

One reason for this widespread influence stems from an issue that frames this investigation: the fact that computing—or computers, or computation[2]—straddles two major spheres of existence. On the one hand, computers are **mechanisms**—effective devices that can be implemented in the physical world, denizens of the causal realm of bumping and shoving, of pushing and pulling, of being electrically or optically or in some other way changeable in physical state, and capable of affecting the physical state of their causal surrounds. On the other hand, computing deals with symbols and information, can produce right or wrong answers, traffics in language and representation. That is, computers play or participate in the realm of **meaning**.

---

information in terms of the amount of structural or configurational order, typically measured in terms of algorithmic complexity. It is these quantitative analyses that are tied into the foundations of physics «…ref Kolmogorov, Adriaans, others…». They also underwrite such claims as that it is information, ultimately, that cannot be transmitted faster than the speed of light. Very different are semantic accounts, which address the question of what information is about in terms of counterfactual correlation (roughly: α carries the information that β just in case, if β were different, α would be different as well) proposed by Dretske and Stamp, and others, and incorporated into the philosophy of mind in terms of teleosemantics «…ref Perry & Israel, Milikan, others? ….».

While the semantic account is not directly formulated in computational terms, Perry and Israel point out that for information defined semantically to be used, it needs to be expressed in some sort of form, which ties into the formal symbol manipulation construal of computing.

[2]Theorists from different perspectives will want to draw sharp distinctions among these words—e.g., by taking 'computation' to be an abstract mathematical notion, and 'computers' to be any physical device that realizes or implements a computation, and so forth. (This last view was famously quipped in Dijkstra's alleged comment that "computer science is no more about computers than astronomy is about telescopes." «…ref?…»). But such distinctions are all *post-theoretic*, in the sense that they derive from one or other specific concept of computation—something to which I am not prepared to commit at this stage. Only once we have a better idea of the layout of the subject matter will it be time to consider how to label its various manifestations.

I use both terms, *meaning* and *mechanism*, in the widest possible sense. By 'meaning' I include the entirety of what philosophers would call the realm of intentionality.[3] Thus I take the term to include far more within its scope than when it is used technically to name something distinct from such other intentional categories as *form*, *reference*, *denotation*, *implication*, *truth*, and so on. The realm of meaning, in this book, includes all of these, as well as *thought*, *belief*, *knowledge*, *speech*, *writing*, *wonder*, and other intentional activities, and such intentional types as *assertion*, *description*, *specification*, *modeling*,[4] *simulation*, *theorizing*, and myriad others.

By 'mechanism', similarly, unless otherwise restricted, I refer to any physical or physically implementable configuration whose constitutive properties, when registered as a mechanism,

[3]In teaching, I playfully tell students that there are three spellings of 'intentional': (i) "intentional with a *d*" (in fact spelled with a 't,' of course), by which I mean the normal use of *intentional* as an adjective derived from the lay term *intend*, as in "Did you intentionally park in the handicapped spot?"; (ii) "intentional with a *t*," meaning its philosophical reading, having to do with reference, aboutness, and various semantical issues (i.e., as having to do with what I am here calling the realm of meaning); and (iii) "intensional with an *s*," also a philosophical and logical term, meaning having to do with concepts, properties, and other notions relevant to understanding and knowledge. Intensional distinctions typically cut more finely than extensional distinctions. The classic differentiating case is Frege's example of "the morning star" and "the evening star": two terms that are extensionally equivalent (both refer to Venus), but intensionally distinct. Someone could know that the morning star is Venus, but not know that the evening star is Venus.

[4]Saying that a model is intentional does not imply that what is modeled need be intentional. Think of a balsa airplane model. Rather, the point is that to call entity a model is to locate it within an intentional realm of modeling. By the same token, to *register*, as I will say, an expression as a *term* (using that word as logicians and philosophers do, as the name for a symbolic structure or expression that refers to or designates an object) does not mean that there may not be other, non-intentional ways of registering it, such as as a string of bits or a sequence of letters. But to describe or use it *as a term*—including using it in order to be intentionally oriented towards its referent, its default use in natural language—is to locate it as an item within an intentional situation of referring.

are expressed in terms of its physical aspects and causal relations to its constituents and impinging causal environment. Mechanisms, that is, as I am using the term, include systems exhibiting emergent behaviour, self-organizing systems, and systems that are not designed (for whom there is no 'mechanic').[5] In particular I do not require that a mechanism be a device designed or used for a purpose or playing a role. Branches fall off trees for mechanical reasons, I would say, having to do with their weight, degree of rot, or perhaps external injury or assault, independent of any intents, goals, or function.

Because my goals is to unearth the ontological and epistemic assumptions underlying common theories, accounts, and characterizations of meaning and mechanism—first in computational systems, then in intentional systems more generally—it is important to be clear, before introducing any more specific physical/causal or intentional vocabulary, exactly what it is being taken to signify. It is to allow for such more specific characterizations later that for now I want to use *meaning* and *mechanism* as broad and largely unrestricted covering terms.

A Two realms

That computing straddles the realms of meaning and mechanism has been evident from the get-go.

That computing has to do with meaning is betrayed in the origins of its technical vocabulary. Some of computing's most basic terms—*symbol*, *data*, *reference*, *value*, *interpretation*, *information*, etc.—were inherited from 17th century discourses

[5] Those who argue that life is not mechanical, to take just one example of a more specific reading, may point to emergent properties, complex systems, etc., as *non-mechanical*. Readers sympathetic to this view will probably do best to assume that by 'mechanism' and 'mechanical,' in this book, I mean roughly 'physical device' and 'physical,' except that by calling a system or device mechanical I mean a physical device *physically registered*—i.e., with focus on its effective physical properties ('effective' because although "being 4.35 light years from Alpha Centauri" is a physical property, it will not count as a *mechanical* property, here, because it is not a property the exemplification of can do local causal work).

about logic and reasoning.[6] These entities are not like *mass*, *force*, *energy*, *charge*, *momentum*, etc.—properties manifestly applicable to moving bodies, physical forces, energetic effects, and such. Rather than stemming from the "empiricist" side of the Cartesian divide, one might say, many of the fundamental notions of computer science derive from the "rationalist" side. They have to do with the kinds of intentional entity studied in logic: expressions and processes that are *about* things—expressions and processes "intentionally directed," as philosophers would put it (and as I will say here[7]) towards entities and phenomena (objects, processes, states of affairs, etc.) that they refer to or are about. The referents will be typically distinct,[8] and often remote, from the expressions and processes that *refer to*, *designate*, or are *about* them.

That computers are mechanisms is equally obvious. What people needed to do, in the ancient history of computation, was to build devices that could be interpreted as computing. The abacus is perhaps the simplest example. When the mechanical arrangement of beads is semantically interpreted according to a prescribed regimen as denoting arithmetic quantities, and then mechanically manipulated in ways that honor that interpretation scheme, the resulting configuration can be interpreted, according to the same scheme, as denoting the results of specific operations on those designated entities. That is, the operations of the device are simultaneously understandable in two distinct ways: first as mechanical changes to the device's physical state; second as abstract operations on the entities denoted by the device's configurational states.[9]

---

[6]«…check with … Güven? Lanier Anderson? … »

[7]I use the phrase "intentionally directed" in this book in part because it is not a familiar phrase in computer science, and therefore avoids the perils of "redefined vocabulary" identified in §«…».

[8]This is not to exclude self-reference, the bane of 20th century theorizing, which has long been a concern of mine. See «…ref CR, Varieties, etc.…».

[9]In the case of the abacus, the mechanical operations involve adjusting the position of beads; the semantic operations, abstract arithmetic operations

This bipartite division of labour, with the gadget or device performing the mechanical functions, and people supplying the semantic interpretation,[10] undergirded the original understanding of computation, and is still widely assumed today.

What rocked the intellectual world in the early decades of the 20th century was the recognition that this model could be extended beyond arithmetic calculation to logical inference. It was amazing, and counterintuitive to most people at the time, that in suitably proscribed domains one could start out with a causal or mechanical configuration of a syntactic representation of one situation or set of facts, let the mechanism proceed, and it would yield up new syntactic configurations that, on the same semantic interpretation scheme as that with which one started, would make sense. And not just make sense, but be semantically warranted. Just as the mechanical configuration of an

---

on numbers. Initially, the mechanical operation of the device was manual, but over time more and more of the motive force of computational devices was replaced by automatic devices—motors, when the manipulation was mechanical, then "motors in the realm of electricity," including actuators, switching relays, and magnetic cores, and ultimately electronic switches made of tubes, transistors, and integrated circuits. That is, the dynamics were originally mechanical (in its narrow sense), then electrical, then electronic. (We think of devices as *electronic*, rather than merely *electric*, when we understand them as mediating the flow of information, not merely of energy). It is that expansion into the realm of information that betrays the characterization's having moved into the realm of the intentional—as involving *meaning*, that is, in the broad sense in which I am using that term.

[10] «…say: called an 'interpretation function' in logical discussions…» By a 'semantic interpretation' I mean a referential or denotational mapping from the mechanical states of the mechanism onto the realm that the device was about—typically a simple arithmetic one for the abacus, and more complex for Babbage's creations. The phrase 'semantic interpretation,' however, has become ambiguous. As mentioned briefly in «§…», below, and explored in more depth in ch. 2, computer science has redefined both 'semantic' and 'interpretation' to refer to mechanical relations and entities, in such a way that, in computational parlance, 'semantic interpretation' refers to relations and entities that remain within (what I am calling) the purely mechanical realm. This is just one of the reasons why we have to tread slowly and carefully with intentional vocabulary.

abacus would denote the result of the arithmetic operation, in a logical system they would preserve truth—leading "purely mechanically," as it were, from true premises to true conclusions.

The wonder was not that these new computational systems *understood* anything. It was evident that it was only when understood in terms of the human-supplied semantical relations to the world that these systems could be understood as computing, performing inference, or doing anything beyond evolving as a physical device. Nevertheless, that a "merely mechanical" device could play a role in seemingly arbitrary intentional processing was an extraordinary and epochal realization, one of the most fundamental intellectual advances of recent centuries.

This bipartite division of labour underwrites Haugeland's memorable characterization of computers as **semantic engines**. Arrange the causal rending of the syntax appropriately, supply a semantic interpretation, start the device on its way, and—*mirabile dictu*—the result will be semantically sensible. All the device had to do was to manipulate the syntax, and "the semantics would take care of itself."[11]

What interests me here is not just the nature of meaning and mechanism, but also relations between the two realms. First I look at the computational case, and then turn to intentional systems more broadly. The unspoken thesis underlying the promise of computing as an idea is that we should be able to apply lessons from computing to the more general case of intentionality as a whole.

B Confusion

One might imagine, given its apparent success, that because computing straddles the two realms, the development and

[11]By "the semantics would take of itself" Haugeland meant that no additional mechanical effort was required to "bring the semantics along" to ensure that the resulting state remained semantically appropriate. This is yet more testament to the non-effective nature of semantic interpretation.

deployment of computational systems, and the accompanying development of computational theory, would have clarified relations between them—perhaps illuminating each from the perspective of the other.

Unfortunately, exactly the opposite has been the case.

Consider just one way in which computing might have borne on the relationship. Given the influence of the reigning scientific/mechanical world view, and in light of the fact that the original task of building computers focused on mechanism, one might imagine that computing or computer science has *mechanized semantics*. That is, as philosophers would put it, computer science might be thought, at least in the computational realm, to have "naturalized intentionality," in the sense of bringing the phenomenon and its explanation within the contours of the natural sciences.[12]

I expect that this assumption would be endorsed by many computationalists, but I believe it is false. Even the most basic results in the theory of computation depend on unexplained intentional relations in the subject matter, which current theory does more to obscure than explicate. Even the relation between binary numerals and numbers[13] results from an act of semantic (intentional) interpretation. Positional encoding had to be invented; the scheme resulted from a creative intentional act. It was by no means discovered as a result in "natural science," in the sense of being a purely physical or mechanical relation.[14]

One reason why the meaning/mechanism relation has been

---

[12]Tellingly, philosophers would be unlikely to say that computer science has "'naturalized computing," since they would presume that computing is already, as it were, a scientifically echt subject matter.

[13]Numbers are neither binary nor decimal; such predicates as *binary* and *decimal* apply only to their expression or representation in a positional scheme.

[14]Arguably, actual cardinalities can be exemplified in the physical world, but the numbers denoted by such elementary expressions as 10^{100} or 50!, or the 256-bit numerals regularly used in online encryption schemes, are too large to be exemplified in the known universe.

obscured derives from the way in which theoretical computer developed. The theory is almost always expressed mathematically, with neither clarity nor agreement as to whether computation is *itself* a mathematical subject matter, or whether, as is standard in other sciences, computation is a (concrete?) phenomenon being *modeled* mathematically. What, to put the issue in question form, is the metaphysical origin of the powers and limitations of effective computability? Do the fundamental computability constraints arise from physical constraints, so that if physics were different, there would be changes to what is and what is not computable? Recent literature conveys the sense that in many or even most cases authors believe that this is so,[15] or at least that this has become the prevailing view, but it is not an issue on which there is general agreement. And note as well the occurrence of the word 'effective' in the theory: it is called the theory of *effective computability*. Can that efficacy be explained as an abstract mathematical property, or is it ultimately physical or concrete efficacy?

Perhaps surprisingly, there is not even agreement on the status of what it is that is computed. By contrast, note the stark difference between the words 'utter' and 'describe.' If I say "I found a great deal on an old Servel refrigerator," I have uttered a ten word sentence, and described an out of date and potentially lethal gas-powered device. I have not *described* a ten word sentence, though I can do so (in fact have just done so) by shifting up to a meta-level, and taking linguistic expressions as my new realm of discourse. And of course I cannot *utter* a refrigerator. But if I say "I computed the value of 13!," it is ambiguous as to whether what I have computed is an abstract number—a number also denotable by (tokens of) the decimal numeral '6,227,020,800'—or whether what was computed was that latter decimal numeral itself (or at least a concrete token of that numeral type).[16]

---

[15]«…need refs…»

[16]If the reader has the reaction that of course one cannot compute an

By themselves, numeric examples may seem trivial. But the issues ramify, and in complex cases involving artificial intelligence (AI), self-reference, reflection, and other more involved constructs, the intentional relations grow more complex, and being clear about semantic or intentional relations is prerequisite not only to theoretical clarity but also to architectural design[17] and to normative and ethical questions about how we should assess the computational results. Do AI vision systems recognize cats, for example, or political suspects, or lung cancers, as their proponents and detractors usually put it? Or are they merely recognising *images* of cats, people, and cancers? Suppose a given system is given a score of 94% accuracy on predator recognition. That presumably means that it has been shown to be accurate on 94% of the tests of images on which it has been tested. But needless to say, its ability to detect the concrete presence of actual predators within camera view, rather than merely correctly classifying 94% of the images on which it was trained (where "correct" means classifying them in a fashion that accords with the judgment of human labelers), depends on the systematicity of the semantic relation between images and actual organisms. No one would doubt that that semantic relation is relevant to the system's real-world accuracy; at issue is whether that semantic relationship is considered part of the computation, or ancillary to it (subsequent, as it were).

Or consider information. Suppose a computational system is used to "distribute information" about forest fires in Quebec, the spread of Ebola in sub-Saharan Africa, or rising interest rates. As noted above, there is debate about what information is—whether, on the one hand, information is itself a semantic commodity, and only represented by formal or linguistic expressions, or whether, on the other hand, information is an issue of mechanical or effective arrangement. The former

---

abstract number, merely in virtue of the fact that it is abstract—i.e., that what can be computed must be material (mechanical)—that betrays a priori allegiance to what I call *causalism*, described below.

[17] «…ref; explain reflection — cf. CR…)

approach follows in the tradition of Dretske, Stamp, and the teleosemantic tradition in the philosophy of mind; the latter is argued by proponents of quantitative accounts of information, such as Adriaans, Kolmogorov, and others aligned with the Shannon-derived "theory of communication" (itself a loaded term).[18] If we do not agree on the fundamentals of information, that is evidence that the foundations of computing, or at least the information processing construal of it, are unclear as well.

Another issue adds to the foundational disarray. Information processing is just one of half a dozen construals of computation. Other conceptions include treating computing as, for example, formal symbol manipulation, the behaviour of discrete (digital) automata, or the realization of effectively computable functions, grounded on models of Turing machines. Famously, these different construals are "proved equivalent," in the sense that it can be demonstrated, given some basic assumptions, that they compute the same (mathematical) functions. But the metric of equivalence used in those proofs is extraordinarily coarse-grained, and the equivalence purely extensional. It is a general norm on theories, however, not merely that they be extensionally correct, but also that they explain and lead to conceptual understanding. That is, they require an intensional characterization of their subject matters. And intensionally, the various construals are distinct. The functional equivalence proof—that, as it is said, the various construals are all "Turing-complete"—and the unanimity buttressed by its uncritical use, is another way in which the foundations are obscured.

Some of the foundational unclarity is manifest in current philosophical debates about computation. In Hall's useful categorization,[19] current philosophical theories of computation can be classified into three groups, according to whether they take computing to be: (i) a purely physical, syntactic or formal phenomenon, or as I would put it, something purely

[18]«…ref to all these people…»
[19]«…forthcoming…»

mechanical, as argued by Piccinini and others;[20] (ii) an abstract or mathematical notion, contingently realized in concrete physical devices (as assumed in what Hall calls "mapping accounts"); or (iii) something intrinsically semantic, in the sense of constitutively involving semantic or referential relations, implying that an adequate theory of computation must rest on (or include) a theory of semantics, as argued for example by Shagrir.[21]

In sum, confusions abound in the intellectual foundations of our understanding of computing. Moreover, most of the unclarity arises from unresolved issues at the meaning/mechanism boundary. That the issues have remained unclear for more than fifty years stands witness to the fact that the development of computational theory has not done very much to illuminate this fundamental dialectic.

C Scientific Revolution

To understand how computing connects the realms of meaning and mechanism, and therefore to assess computing's ultimate promise as an intellectual idea, we need to step back and take stock of our current understanding of these realms in sufficient depth to underpin an analysis of what computing is, and how it pertains to their relation. To do this, and understand the intellectual context of computing and computational theory, it will help to go back to the origins of the Scientific Revolution in the 16th and 17th centuries—to the rise of the natural sciences, and to the emergence of the "age of mechanism."

The scientific worldview ushered in by such pioneers as Copernicus, Galileo, Bacon, Hobbes, and Newton was founded on two interlocking norms.

Epistemically, it wrested authority away from church, state, and divine revelation. Allegiance was pledged neither to the words of others, nor to the authority of sages, nor to other-

[20] «…refs…»
[21] «…refs…»

worldly entities. Deference was instead granted to the "world itself"—the world towards which our words and thoughts are directed, the world as revealed through experiment and direct experience.

Ontologically, an overriding norm was to avoid reference to anything "spooky"—preëstablished harmony, élan vital, and the like—unless the existence of such entities or phenomena could be empirically verified. While it is impossible to provide anything but a vaguely circular characterization of what it is to be 'spooky,' coherence with direct empirical evidence was perhaps the most critical factor.[22]

I take these two norms—granting authority to the world itself, and avoiding what is spooky—to constitute the essence of science's claim about **nature**. Natural science, on this reading—and as I will myself will use the term[23]—is thus the study of the *nature* of the world, and **naturalism** an appropriate name for that study's associated methods.

C.1 Causalism

The concepts of deference to the empirical world and the

[22]What bars extra-sensory experience from serious scientific attention is not only the fact that it involves reception of information not gained through the recognized physical senses, or the fact that it violates the "no action at a distance" norm which seems so overwhelmingly to govern concrete physical operations and phenomena (at least those at our mesoscale human level). These are challenges enough, which defy obvious solution. More seriously, it is not just that claims of extra-sensory perception do not seem to be true; it is also that we have *no idea what it could mean for them to be true*. Evidence on their behalf would require a complete overhaul of our fundamental understanding—an understanding that undergirds our entire current grasp of the world.

[23]The term *science*, and subsidiary terms such as *physics*, *chemistry*, *biology*, etc., are sometimes used ambiguously both as the name of a human epistemic project of understanding the natural world, and for the content of the world so understood (as in "the chemistry of all known life depends on carbon bonds"). The former meaning most accords with its etymological roots in the Latin *scio* (to know), and is how I myself will primarily use these scientific terms.

rejection of anything not rooted in it do not exhaust the character of the natural sciences, however. A third ingredient has been constitutive of science since the very beginning: its focus on the *physical* world, and on the causal laws that govern it. Determining the nature of the causal laws has been the primary aim of the scientific project to date. Equally important to contemporary science, if less frequently noted, is the assumption that physical phenomena[24] (objects, processes, etc.) are what they are in virtue of the way they instantiate causal properties—the properties in virtue of which the governing laws obtain. Objects in particular are assumed to be causally constituted and causally individuated: common causes of multiple effects, common effects of multiple causes.[25] Causal properties, at least within scientific realms, determine objects' identity—makes them what they are, distinguishes one from another, differentiates one from two, determines whether they have remained the same or changed over temporal interval or in the face of some "external" event, and so on. Interaction is understood to result from the impinging of causal forces on phenomena's physical states.

Because of the nature of all physical laws, moreover, states and interactions are understood to be governed by a form of space-time localism. Causal forces, at least at the mesoscale level relevant to human experience and computation as we know it, appear to act locally.[26] Physical states and interactions result from proximally impinging stimuli or forces occurring, roughly speaking, within a $1/r^2$ space-time envelope. No action at a distance, either temporal or spatial. These constraints are

---

[24]I freely use the term *phenomenon* to refer to entities, processes, and other ontological posits, without intending anything "phenomenal," in the sense of being a subjective property of human observation or experience.

[25]«…Campbell? is he the originator of this framing? Ask Güven…»

[26]«…talk about quantum entanglement and consequent non-localism; doesn't change my claim …… Point out that causation is relativistically consistent, too: if α precedes β in any reference frame, it will precede β in any other relativistically consistent reference frame … »

embedded in the widespread use of differential equations to express scientific regularities.[27]

The assumption that the world is fundamentally causal is a third property distinctive of contemporary science—additional to granting authority to the world itself and avoiding what is spooky. It can be codified in two somewhat more specific ways. The first is **causal closure**: nothing can be the cause of a physical event, nor can it be caused by a physical event, except other physical objects, processes, and events. This closure condition not only rules out divine intervention; it also blocks the existence of what is symbolized by Descartes's pineal gland: any entity that mediates influences into and out of the causal realm (e.g., to or from spirit or abstract mind). The second more specific claim is **causal completeness**: nothing in the world exists (with the possible exception of mathematics and logic) that does not supervene on the totality of the causal plenum.

Philosophically, the closure and completeness claims are understood in terms of what is known as *global physical supervenience*:[28] the thesis that there can be no change or difference in the world at a "higher" level of abstraction or idealization without there being some (subvening) change or difference at a "lower" level of abstraction, ultimately in the fundamental physics.[29] If in one world I know or think something that I do not know or think in the other world, then there must be a difference in the underlying physical character of the two worlds—a difference on which the mental difference supervenes.

I will take the union of two these two theses, causal closure and causal completeness, to constitute **physicalism**—a metaphysical worldview from which nothing I say in the present investigation will diverge.

This trio of commitments—metaphysically to an overarching

[27]«…Ref 03…»
[28]See the sidebar on reduction and supervenience on the next page.
[29]Relate to what Haugeland' calls "weak supervenience" «…ref…».

physicalism, ontologically to global supervenience, and epistemically to naturalism as the appropriate method for its study—has undergirded natural science for centuries. and

Reduction and Supervenience

Both *reduction* and *supervenience* have to do the relation between a situation or phenomenon, or even the whole world, registered at a one level of abstraction, and that same situation, phenomenon, or world registered at a lower or more basic level of abstraction. Often, though not always, the lower level of abstraction is fundamental physics, on the grounds of its being the lowest level of abstraction that exists, or anyway that we can and have registered.*

Given that broad similarity, *reduction* is taken to be an epistemic relation between concepts or theories constitutive of our understanding, in contrast to supervenience, which is understood as an ontological or metaphysical relation between the world or part thereof, as registered at the higher level of abstraction, and that same world or part registered at the lower or more basic level (or between the upper and lower sets of properties).

Classic examples of reduction include water understood as H_2O, and temperature understood as mean molecular velocity. Water and heat are the higher-level concepts; hydrogen, oxygen, and molecular velocity are lower-level, being notions from physics or chemistry. The relation is reductive because it relates the concepts in terms of which we understand the phenomenon in question. This leads to intelligibility of the claim that water simply *is* H_2O, and that temperature *is* mean molecular velocity. If concept α can be reduced to β, where β is intelligible is a notion of the lower-level theory, then it is intelligible (and common) to *define* α as β. Thus we might say that the claim that temperature is mean molecular velocity is analytic, because that is how temperature is defined.

Supervenience is trickier to characterize. The standard definition is that there can be no difference between the two situations or worlds in respect of the higher-level of abstraction, without a corresponding difference in that situation or world at the lower level of abstraction. More compactly, if γ is an phenomenon at the higher level of abstraction (i.e., exhibits various higher-level properties), and δ is the lower-level phenomenon on which γ supervenes, then there can be no change or difference in γ with-

continues to hold sway today. Any properly "scientific" theory, according to the nearly universal consensus, must be an account of causally-individuated phenomena (entities, processes, etc.)

out corresponding changes or differences in δ.

Reduction implies supervenience, but not the other way around. Thus many people (all materialists, among others) will believe that social arrangements, and phenomenon such as trust, authority, deference, etc., ultimately supervene on various stupefyingly complex arrangements of underlying physical molecules, but no one would expect that we will ever have an intelligible account of just which physical arrangements of molecules constitute a social situation of trust, and which ones do not. Similarly, it seems safe to assume that tokens of currency (quarters, dollar bills, paper rupees, etc.) are all physical objects, and so, individually, have at least relatively stable underlying physical constitution. But at the level of fundamental physics, what it is to be a quarter, a dollar, a rupee is presumably of an order of complexity transcending coherence or intelligibility.

I will distinguish what I will call **local supervenience**, when the upper-level phenomenon γ supervenes on the exemplification of lower-level properties within the region of space-time occupied by γ, and **global supervenience**, in which γ supervenes on the total state of the world at the lower level of abstraction (on the exemplification of lower-level properties by the entire world), but need not supervene on the exemplification of lower-level properties within the region of space-time occupied by γ. For example, suppose γ is the property of "being owned by emperor." An object exemplifying γ—a statue, say, or a chariot—might occupy a proscribed spatio-temporal region ε, but even if the exemplification of γ ultimately supervenes on the entire world's exemplification of fundamental physical properties, it would likely not supervene on the exemplification of fundamental physical properties by ε alone, since it would depend on facts about the emperor, not just facts about the statue's or chariot's physical character.

*In the philosophical literature, the levels of abstraction relevant to reduction and supervenience are typically characterized ontologically, in terms of sets of properties. Registration, in the sense I use it here, bridges the epistemic/ontological boundary. See O_3.

exhibiting causal properties, and involved in causal interactions, unfolding in accord with locally governing causal laws or regularities.

This makes science as it is currently practiced into something like a 21st century update of the 17th century philosophy of mechanism. Needless to say, radical advances have been achieved in our understanding in the intervening centuries: not just of specific physical phenomena, such as relativity and quantum mechanics, but of diverse types—emergent phenomena, non-linear dynamic regularities, self-organized systems, etc. As I have indicated, what it is to be a mechanism needs to be substantially broadened from what was assumed four centuries years ago. Yet the causal spirit of the movement endures. It is the view widely represented as being the correct rendering of *naturalism*, and theories framed in its terms as exemplifying the proper form for *naturalistic theories*. Because I want to argue for a broader reading of these last terms, however, I will refer to this view, the currently prevailing view of science, as **causalism**.[30]

C.2 Influence

The influence of causalism has been immense. Proposals that violate its strictures are typically derided as unscientific. Moreover, this normative influence has spread much more widely than simply within the "natural sciences" where it was first

[30]I avoid the term 'materialism' for several reasons. First, I believe that 'materialism' is most often used as a name for a metaphysical or ontological position: that everything that exists, with the standard exception of mathematics and logic, rests on physical foundations—in essence the view for which I am using the term 'physicalism.' Second, 'material' can carry normative connotations and have normative uses, as in "that is not a material consideration." Many readers will say that that is simply a separate and different meaning of the term—but I do not agree. In 03 I argue for understanding a 'material object' to be a *patch of reality that matters*, not merely an uninterpreted chunk of the physical realm.

By 'causalism' I mean to signify not just a metaphysical/ontological position, but commitment to a form of its explanation.

developed (physics, chemistry, astronomy, and perhaps biology). Causal properties are today often tacitly assumed to underwrite the individuation of all entities and the regularities that govern them. As well as governing their identity, they are used to make the distinction between *intrinsic* and *extrinsic* (or relational) properties.[31] Non-causal properties—being the object of someone's desire, being worth $100, being of average height for a 14-year-old, and the like—are automatically characterized as relational or extrinsic.

Overall, the causalist world is our contemporary understanding of how to understand Descartes's *res extensa*.

Extraordinary effort has been devoted in recent decades to developing "properly scientific"—which is to say causalist—accounts of phenomena that, at least on initial reflection, are not obviously constitutively causal.

Three examples. The first is a widely held doctrine, known as f*unctionalism* in the philosophy of mind, which underlies the entire project of artificial intelligence and at least arguably computer science as well: the idea that our mentality, and intelligence more generally, is constituted by patterns of organisation (neural in the human case) playing distinctive functional roles within a causally bounded area (roughly, in the human case, the brain, though in theories of embodied cognition this may be extended to the full nervous system or body). If those same patterns and processes are realized in other material substrates, the idea goes—silicon, for example—the results would also be authentic minds, or authentic intelligence, just differently implemented. By the same token, if a brain implant were devised that played exactly the same functional role as a part it replaced, the resulting amalgam would not only be a genuine mind, but in fact would be the *same* mind.

[31]Lewis's characterization of this distinction is wholly framed in terms of an unexplicated presumption of object identity, universally understood in scientific contexts to be causally determined. «…ref "Extrinsic Properties"…»

Crucially, the idea of a "function" underlying functionalism is restricted to a causal function—that is: a function that can be played in virtue of a state's causal or effective profile—on the model, it is assumed, of a similarly causal state in (a physical realization of) a Turing machine or other computer. Characterizations of Turing machine states rarely if ever mention explicitly that they are constrained to causal individuation, but the condition operates as a tacit background "well-formedness" condition. To violate it—e.g., by declaring a controller to be in state σ_1 if the continuum hypothesis is true, or if Kurt Gödel ever thought about that state, or if it has ever before been physically implemented—would be to cheat.

While it is difficult to formalize the well-formedness condition non-circularly, the condition is intuitively clear enough. To put it in blunt if informal language, different Turing machine states must be such that a physical realization can be constructed (if Turing machines are not taken to be physical to start with) such that if they are in one state, they will turn on an electrical switch, and in another, will turn it off.

A second and by now equally accepted example of efforts to "causalise" what is not obviously causal on initial reflection comes from biology, and its attempt to causalise normativity in terms of evolutionary flourishing. It is assumed, first, that concepts of biological function can be cached out, on such an approach, in terms of the (causally-individuated) role they have played in ensuring or contributing to the species' overall (causal) evolutionary fitness. Thus the function of the heart can be causalised as pumping blood, and the function of a sperm as fertilizing eggs, because, over time, it is those activities that have led to the evolutionary survival of the organisms in which they occur, even if the statistical frequency of those activities is low (as is dramatically the case for sperm). The normative—the "correct" or "good"—can then be defined in terms of that function which maximally promotes that evolutionary flourishing. Given such an account of function, tools then become available to invoke such normative characterisations as that "the heart is misfunctioning" (rendered along the lines of its being

maladaptive) without leaving the causalist boundaries of scientific explanation.

A third example, building on the second, shows how far the causalist project has been extended beyond its origins. That example is evolutionary psychology—the effort to bring the realm of thinking, reasoning, language, thought, and the like into the causalist program. Just as the heart evolved to pump blood, and the liver to detoxify poisons, so too psychological states such as claiming and referring are argued to have evolved to solve environmental problems.[32] And in the same vein, evolutionary approaches are applied to such issues as altruism, fidelity, benevolence, etc., in an overall effort to derive "ought" from "is."

What is perhaps most striking about these efforts, in a crucial step against which I will presently argue, is their assumption that the causal story of *how a capability arose* and of *how it functions in evolutionarily adaptive ways* is the appropriate and complete scientific account of *what that capability is*. "Is," that is, is reduced to how it arose, how it got here, and what causal function it plays.

Is this sufficient? I do not think so. It makes it impossible to claim that humans, and perhaps other organisms, "discovered" (that is: evolutionarily stumbled upon, and evolved so as to exploit the fact) that referring to objects and making true claims was evolutionarily adaptive. The reason that such a claim cannot be meaningfully advanced within the program of evolutionary psychology is that referring to objects and making true claims are effectively defined as *that evolutionarily adaptive activity that humans* (and perhaps other organisms) *evolved in order to solve various problems of fitness and survival*. In this way the causalist program turns the thesis that referring to objects and making true claims is evolutionarily adaptive into a

[32]« … Ref Milikan (especially her book "Language, Thought, and *Other Biological Categories*: New Foundations for *Realism*" [emphasis added]), Dretske, others?…maybe Searle?…»

tautology—thereby evacuating it of meaning. And the program does not stop at thought. Such personal-level[33] epistemological normative states as altruism, generosity, fidelity, and the like, are, like thought and language, corralled into the causalist paradigm, via grounding in evolutionary flourishing. By fiat, the train on which a capability arrived and how it aids evolution is taken to be the account of what capability the train delivered. Full stop.

It will be clear, when we get to intentionality and computation, that there is more to explaining and understanding intentional phenomena than providing a causal account of how they arose, how they are implemented, and what causal consequences they engender. There is nothing especially rare about this structure. There is more to understanding why there are just five Platonic solids than is contained in the content of a causal recipe for constructing them. Purely syntactic accounts of logical inference are intrinsically incomplete without an account of what the expressions being manipulated mean or designate (*sans* substantive notions of intentional content, the best one can do is to reduce soundness to proof-theoretic completeness or another overall syntactic relationality). So too there is more to intentional and computational phenomena than accounts of the causal operation of their mechanical implementation. Part of the reason is that the non-effective nature of being intentionally directed towards some potentially distal part of the world is an essential aspect of reference and intentionality, implying that no purely causalist explanation can do justice to the semantic and the intentional—can fully explain Haugeland's semantic engines.

Evolutionary psychology is not the end of it. Another example of how far the project of developing causalist accounts of natural phenomena has traveled from its origins in the lower-level

[33] «…explain personal vs. subpersonal…cite McDowell's "Content of Perceptual Experience"…»

natural sciences is contemporary cognitive science.[34] The past few decades have seen an upsurge in theories of "embodied cognition," which eschew referential, representational, and other intentional accounts of mind in favour of embodied, embedded regimens of causal (including bodily) activity.[35] In contrast to grappling with the challenges that referential and representational phenomena raise for purely causal explication—one of my aims in this book—the embodied cognition movement not only sees cognition as necessarily embedded and embodied, (with which I agree), but as sufficiently explained thereby (with which I do not).

Note that the allegiance that cognitive science pledges to this causalist project is epistemic as well as ontological. The adherents are not merely attempting to anchor cognition within Descartes' res extensa, or endorsing physicalism. They are striving to incorporate its cognition's *full explanation* into a causalist framework as well—into the epistemic framework first developed for forces, masses, and manifestly purely causal phenomena.

Causalism also underlies an increasing number of theoretical efforts in the humanities and social sciences. Consider for example the "materialism turn" (and "new materialism) adumbrated in feminist epistemology, science and technology studies (STS), and critical social theory.[36] Think too of Barad's endorsement of "causal intra-action" as the basis of a radically non-reductionist ontological picture. While Barad has a much wider conception of causality than is traditional,[37] she too retains allegiance to a causalist explanation of the world.

---

[34]That cognitive science has always remained within the causalist paradigm is reflected in its retention of the term *science* in its label.

[35]«…Ref Chemero, others?…»"

[36]«…say more about these things? At a minimum need refs…»

[37]I believe her project is something like this: to show how an appropriate but radical conception of what she calls *causal intra-action* can provide a subvenient basis for everything concrete or occurrent, presumably including mind.«…check that this is right…»

Many of these causalist projects in the humanities and social sciences are defended, either explicitly or implicitly, as thoroughly non- or even anti-reductionist. Their advocates do not expect the phenomena and regularities they posit to be derivable, even indirectly, from the underlying "physics" properties of their material realisations (their subvenient basis). This rejection of reduction is sometimes taken as evidence of their freedom from what are viewed as the untenable strictures of the scientific paradigm. Yet I believe this anti-reductionism is ultimately somewhat superficial, or at the very least partial, glossing over a deeper and enduring privileging of causal forces and relations. The theoretical emphasis currently being placed on material embodiment and its local activities betrays this emphasis.[38]

Reductionism may have been buried, or at least submerged. But causalism still reigns supreme.

C.3 Intentionality

What then about intentional phenomena? Has the causalist zeitgeist affected the intellectual project of understanding intentionality and intentional processes? How has the effort to bring wider and wider swaths of intellectual terrain within the compass of an ever-expanding causalist scientific worldview dealt with the challenges of thought, meaning, language, symbols, reference, and truth?

Reference, truth, and general intentional "aboutness" pose an obvious challenge to causalism, since they are neither locally confined nor self-evidently causally individuated. Some of reference's non-local aspects have been highlighted in what in philosophy are called "externalist" theories of content. If I say "I hear that fermions have half-odd-integer spins: 1/2, 3/2, 5/2, etc.", what I refer to by the term 'fermion' is determined by the community of physicists to whom I defer, not by anything I myself know or believe. Similarly, my thought about fermions

[38] «…get quotes from Suchman? Wilson? Clark?…»

may differ from my thought about bosons not in virtue of anything (causally) different within my skull, but in virtue of the referential language practices of experts within communities of which I am a part. Nor does the lack of any internal causal difference inside me between the two cases imply that there is not a fact about what I am referring to; reference is secured not only by my membership in the relevant linguistic community, but also by the nature of the world I thereby refer to.

By the same token, reference is not *effective*. Referring to something does not bathe its referent with causal energy or force. No matter how successful, a reference to an entity is not something that, in and of itself, qua successful reference, can materially affect the referent. Because of this, there can be no such thing as a "reference detector"—a device which would fire whenever the object to which it is attached was the object of someone's thought or utterance.[39]

Yet none of this means that reference is not *concrete*, in the sense of being spatio-temporally occurrent. Nor can its *actuality* be doubted. Without reference thought would be impossible; we would not exist as us; in fact a reference-free existence is by definition unimaginable.[40] Reference is also *experienced*—as for example if you take umbrage upon being referred to derogatorily. Any objection that you do not experience reference itself, but only the immediate causal profile of the referring act, stems from a prior belief that one cannot experience that which is not causally efficacious—the very point under contention. That is not to say that there is not something profoundly right about the implicit deference to causal completeness. Understanding causal efficacy's role in intentional behaviour is of the utmost importance, and must be explained by an candidate theory of intentionality. And to repeat a point made above, nothing I am

[39]See *Promise*, ch. 2, fn. 14, where I describe an iPhone program that would beep whenever anyone in the world thinks about its bearer. Needless to say, the program cannot be built.
«…explain impossible iPhone app—where do I talk about that?…»
[40]Reference is as secure as cognition in Descartes's *cogito*.

proposing is a challenge to an overarching physicalism. But, crucially, the causal envelope does not encompass the whole discursive phenomenon. What one takes umbrage at, if insulted, is a reference, and its attribution of character defect or malign properties, not the causal profile of that reference's exemplification.[41]

Just as I see a person, when I gaze across the street, not the laminar incidence of a pattern of spectral illumination that is providing me with the information that they are there, so too reference directs my intentional attention to a distal referent, not to anything proximal, not to any effective proxy or enabling material condition. To confuse the two is to conflate the subpersonal with the personal.[42] What I experience, when a long-lost friend walks into the room, is the *friend*, not the spectral configuration of visible light that my brain used to determine their identity. When I hear an invitation to lunch, a recommendation that I go home, or a claim that a colleague has landed in Delhi, it is towards lunch, home, or Delhi that my intentional state is directed, not the causal profile of the stimulus that caused me to be so oriented.

The most significant long-term intellectual impact of Descartes's dualism, in my view, was neither his argument for the existence of God, nor his separation of a concrete realm (*res extensa*) from an abstract one (*res cogitans*), with cognition and intentionality located in the latter. Rather, it was the division of labour thereby effected. The world of physical forces and material[43] objects—bumping and shoving, local constitution and liminal impact, inherent vs. relational properties, etc.—was

---

[41]It is the reference, not the form of the expression, that matters. If you discover that the term you thought denoted you was in point of fact directed at someone else, your discomfort would evaporate. What matters to you is whether you were referred to, not how you were referred to.

[42]«…explain; and reference McDowell "The Content of Perceptual Experience"…»

[43]«…note on 'material;; cf. O_3…»

separated out as a world to be investigated on its own, with its own autonomy, freed from inexorably complicating issues involving time- and distance-spanning thought, reference, language, theory, imagination, and, at least on the surface, experience.[44] Purged of intentional complications, the causal world took pride of place for empirical inquiry. "Natural science" proceeded on its causalist sojourn, unfettered by the complications of non-effective, long-distance intentional relations.

The unparalleled success of the causalist program over the next several hundred years has gradually led, in spite of Descartes's split, to a myriad attempts to re-unify the intellectual realm, but in an asymmetrical way: by corralling the intentional phenomena into the causalist paradigm. From Aristotle's articulation of modes of inference, to Boole's and Peircian logical investigations, to the development of formal logic and computing, one sees inchoate attempts to causalise rational behaviour, ultimately leading to the logical inference machines of the first part of the 20th century, in which the formal (syntactic) aspects of logic were mechanically implemented in concrete physical devices.

The logical story is more complex than simply a recital of its formalist projection, though. As suggested above, and explained in more detail in chapter 2,[45] a proper understanding of formal logic involves recognizing the central roles of two distinct relationalities, not just one: a non-effective (i.e., non-mechanizable, which is to say not necessarily explicable in causalist terms) semantic relationality in addition to a causally mechanizable syntactic one.[46] Only by first distinguishing and

[44]«…pace the fact that we know of mass, weight, movement, etc., though *experience*; but this role could be chalked up as part of the epistemic apparatus of doing science, not to be worried about as constitutive of the subject matter being investigated…»

[45]«…where? also CR?…»

[46]That syntactic relationality be mechanizable is a tacit well-formedness condition, with the caveat that syntax is sometimes modeled mathematically, but that mathematicization must still be of a mechanizable relation.

then relating the two realms is it possible to define substantial notions of soundness, completeness—and truth. Nevertheless, even in spite of what might seem an obvious and elemental point, the pressure of the causalist program has led to efforts to fit the entire enterprise into causal dress, with the introduction of term models,[47] definitions of completeness in terms of all possible syntactic derivations, etc.

Initiatives to bring intentionality into a causalist framework are not limited to logic. Philosophers of mind have introduced notions of "narrow" content, taken not only to be causally interior to the brain, but to consist of causal organization and causal impact, in distinction to "wide" content, to which external world-directedness is relegated.[48] Think too of functionalist theories of mind (where, as noted above, 'function' is shorthand for causal function), and psychological theories of meaning which locate meaning as a (causally) internal property of brain states. Inferentialist accounts of rational thought, of the sort proposed by Brandom, can be similarly construed: they model rationality on the giving and taking of reasons, which causalist-minded theorists can affiliate, if not identify, with causal features of thought structures exemplified in the brain.

Some of the most striking evidence of causalism's influence in evolutionary biology was mentioned above: the project of evolutionary psychology, and attendant efforts not only to "naturalize" normative aspects of biological life, including biological

---

Although I might define a language with "undecidable" syntax, to do so would be theoretically awry.

[47] I consider the use of term models to theorize semantics to be something of an intellectual cheat—a view shared by the late Jon Barwise.

[48] «…I believe the narrow/wide distinction is ill-formed, as detailed in CR, when, as is usual, it is implied that narrow content (content that does not depend on the subject's environment) can be associated with structures that are local and effective. Thought is permeated with reference to internal states ("I am confused"). Even in case when those internal states are causally efficacious and within causal reach, that does not mean that the referential relation of our meta-level thoughts to them is itself a causal relation…»

versions of function, worth, etc., but more broadly to bring personal-level epistemological normative states, such as altruism, generosity, fidelity, and the like, into the causalist paradigm, ultimately grounded on evolutionary flourishing. And artificial intelligence, to take a very different example, assumes causalism almost as a condition of sense. "Meaning," to a contemporary AI researcher, is widely assumed to refer to causally effective internal (brain or computer) structures. "Meaning representations," to my mind an oxymoronic phrasing, are discussed without a moment's hesitation.[49]

It would take an entire book to present a full critique of the causalist project. Yet I believe it is fair to say, in spite of a panoply of efforts, that none of these causalist projects have come to grips with what I will call have been calling **intentional reach**: the fact that our thoughts and language are non-effectively directed across and beyond the limits of causal connectedness, beyond the realm of causal efficacy, to encompass the whole world (and worlds beyond the actual). Even causal theories of reference, perhaps the most concentrated attempt to naturalize semantics, relies at its base on incidents of "dubbing," where a term is introduced *to refer to the object or phenomenon in plain view in front of the subject*—a construction that takes as given both reference itself, and the world towards which that reference is non-effectively directed.[50]

Does all this mean that naturalism is doomed? No. That is not to say that I know how reference works. I have no story of how sentences "grab onto" the situations they are about, so as to determine their truth or falsity, nor any story about the deference we accord to those situations, to allow them to have normative grip on us. I have no account of how our thoughts and imagination lock onto the distal situations to which we refer. But just because reference is such a basic and natural part of our

[49]«…explain that I would assume the data structures *have meaning*; not that they *are* meanings…»
[50]«…check with Jessie…»

experience that it is in fact impossible to imagine its not holding (imagination would not be imagination, would not be about anything, if reference did not hold), that is no reason to suspect that in and of itself reference does not ultimately supervene on the configuration of the physical plenum in toto.

What I believe the manifest non-efficacy of reference and intentional reach calls for, in the absence of theory (and, again, I will offer no such theory here), is that, as a first step, we understand the non-effective regularities constitutive of intentional phenomena, and how those regularities relate to the causal regularities governing the mechanical operation of those phenomena. That is, we need to treat reference as a first-class regularity, for example, not hobbled or disappeared by causalist bias. Then, if we are motivated by physicalism, we should understand how tokens of a materially embodied species can implement a way of referring. But as computer scientists know in their bones, implementing is not reducing. Even if the design of an abacus is motivated by an understanding of arithmetic, instructions on how to construct one is not a theory of arithmetic. An account of an implementation is not an account or explanation of that which is implemented (unless one has pledged allegiance to causalism in advance).

Earlier I took two tenets to be constitutive of a naturalistic world view: the deferential granting of authority to the world of inquiry and experience, and the avoidance of that which is "spooky." As I will argue below, neither of the these two tenets, on their own or together, imply causalism. There is room for something wider.

C.4 Computation

What then about computing?

The causalist influence on computer science has been extreme. As I will show, by a combination of avoidance, confusion, and redefinition of technical terms, computer science has completely subjugated its analyses under a causalist program.

There are three standard positions on computing that might be thought to relieve us of having to take the intentional

seriously. On the one hand are those who deny that computation is intentional in any way, and therefore believe that all talk of reference and non-effective directness and the like are beside the point. A second group are those who argue that, sure enough, many (perhaps even most) of the computational systems we construct and deploy are intentionally interpreted, and would grant in addition that representation and intentionality are genuine phenomena that will ultimately need explanation, but nevertheless claim that the success of computer science demonstrates that intentionality is inessential to computing per se—inessential to computing *as computing*. For evidence, they will point to automata theory, mathematical theories of computability and complexity, and the like, on the assumption that these do not rely on any intentional notions.[51] The third group will admit that, sure enough, computation necessarily involves representation or information or symbols or some such intentional fare, but believe that these can be reduced to physical properties of the computational system. Whereas the first group deny that intentionality is relevant, the second believe that while it is often pertinent it can be set aside from the core computational phenomenon, and the third believe that we can reduce (at least computational) intentionality to physical or other intentionality-free computational states.

I believe that all three of these tacks fail.

First, as already suggested, the idea that mathematical theories of computing are intentionality innocent is simply false. As will be shown in detail in ch. 2, if computation were not intentional, there is no way that one could describe a (causal) physical process as *computing* anything—adding numbers, determining consistency, figuring anything out. Sure enough, automata theory is not evidently intentional, at least on the face of it. Considered as such, the causal (mechanical) behaviour of automata may not be intrinsically or in any other way

[51] «…Ref Piccinini as the most obvious, but others too…get refs from Jessie…»

intentional. Automata *per se*, however, are merely self-propelled discrete devices that automatically follow a predetermined sequence of steps. Automata theory takes an additional step, however: it characterises the automata it describes as *computing* devices. It is their interpretation as computational that requires the additional step—not just of interpreting them, which any theory does, but interpreting them *as intentional*.

Second, the intentionality gets buried in the abstract character of the (accepted) theory of computation. The problem is not that the theory is expressed mathematically. That much is true of physics, with no implication from the (non-effective) use of mathematical structures to model concrete situations (or situation types) that the entities thereby modeled are themselves abstract—magnetic fields, quarks, black holes, etc. Rather, the difficulty in the computational case lies in the fact that computations themselves, whether thought to be modeled or not, are portrayed as purely abstract mathematical phenomena.

But the mathematical character of the description—the fact that the accepted theory of computation characterises computation as mathematical—does not mean that concrete issues of mechanism are theoretically irrelevant. For an almost trivial example, note that students of computational theory must learn that complexity results are expressed in terms of the length of their standard *numerals* (i.e., in terms of the size of their mechanical encodings), not the not the actual value of the numbers thereby denoted. Thus the number $2^{2^{10}}$ would trivially be taken to be computable (yielding a 1025 bit long binary numeral), even if the cardinality it represents cannot be causally exemplified in our universe. The relation of numerals to numbers is an intentional relation; what warrants it is not explicated in the theory.

Third, there is confusion in the theory about what is sign (model, classification scheme, etc.) and what is signified (the computational phenomenon itself)—and an allied slippage between what is part of the theory or epistemic apparatus used to classify the phenomenon in question, vs. what is proper to the subject matter (computation) itself. Some of these confusions

have been mentioned: the ambiguity as to whether computing is an abstract mathematical phenomenon (famously quipped in Dijkstra mantra: that computers are no more relevant to computing than telescopes are to astronomy[52]), or is in point of fact concrete."[53]

These confusions, and the failure to deal with the semantical interpretation of computational states, may have historical roots in the fact, mentioned earlier, that the first computers, and such precursors as the abacus, were originally conceived as purely mechanical devices we use to compute. That practice may have inexplicitly supported the idea that a theory of them would therefore be responsible only for being a (perhaps relatively abstract) account of them as mechanical devices, not an account of the computational phenomenon they enabled. This would naturally fit into the conception of computer science as a theory of automata or abstract machines, with "what the automata computes" left off the table. But if their computational prowess is ignored, then what results has no claim to being a theory of computation.

C.5 Redefinition

These is a fourth source of confusion in the foundations—a somewhat perverse development that almost wholly blinds computer science to the existence of problems.

I noted above that much of computer science's original technical jargon has intentional roots (*symbol*, *information*, *meaning*, *semantics*, *values*, *reference*, etc.). In spite of that intentional history, something remarkable has happened in the intervening centuries:

> *Within computer science, every classic intentional notion has been redefined—construed as designating something mechanical and effective within the locally proscribed computational system.*

[52] «…ref fn ???…»
[53] «…ref; only alleged?…»

Thus the "semantic value" of a computational variable, to take just one example, is taken to be the (mechanical) computational entity to which it is "bound," or the contents of a memory cell. The value of CURRENT-TEMPERATURE in a weather program, for example, is likely to be a representation of an integer—i.e., a numeral—not a worldly temperature, or even a number;[54] the value of CURRENT-USER, a computational identifier, not a live human being; entries in the "country" field of a database of international students, system-internal computational constants that we humans interpret as naming countries, not the countries so named. And so on.

Similarly, the "semantics" of an operation in a programming language is the behaviour to which its execution leads. In and of itself that may not be intentionally awry, even on the traditional notion of semantics, if, as seems plausible,[55] programs are understood to be meta-level structures "about" operations on data structures and the like. But in a program, the semantics of an expression such as PROFESSOR.COURSE-107 would be described as "obtaining the value of the PROFESSOR entry in the computational object that is the value of the of COURSE-107—again, a canonical representation of a person, not a real-world human being.

This casualization of semantical terminology within computational discourse has been so thoroughgoing that it has grown extraordinarily difficult to express, to a practicing

---

[54]Theories of the semantics of programming languages (especially "denotational semantic" accounts) often describe the value of numerical variable as numbers rather than numerals, but it is easy to demonstrate that that practice is in fact one of mathematically modeling a numeral, not of actually taking the abstract number itself to be the computational value. For example, it is standard for tests of numeric equality (VAR==7, e.g.) to be viewed as perfectly legitimate computational operations. Abstract numbers themselves cannot be directly and mechanically compared, however; what the acceptance of the operation really signifies is that tests of the identity of *numerals* are legitimate and causally implementable (so long as the numerals are represented in as common positional notation scheme).

[55]«…ref CR…»

computer scientist, what the fundamental properties of intentional systems are. The problem is that *computer science has lost the words*. I have sometimes resorted to talking about "the semantics of the semantics of a variable," to get at (for the examples above) the actual temperature, the living and breathing person currently using the computer, the country of Zimbabwe, etc. Even this strategy typically fails, though, because the terms 'semantics' and 'reference' have been beggared into forms of causal link, with the result that genuine reference no longer lies within the conceptual space of possibility.

The fact that computer science has lost the words to talk about intentionality does not mean that we can ignore the subject. We cannot simply observe that within computer science these once-intentional terms have taken on the technical function of referring to causal relations, treating this as a normal example of a scientific field adopting disciplinarily specific meanings of familiar fnatural language terminology for its technical terms—and then leave it at that. This for at least two reasons.

First, the concepts that the terms classically referred to, and what they continue to refer to in disciplines unaffected by computational ideas, are not just relevant to present-day computing, but as I argue here, are *constitutive of computing as computing*. When I say that computing is essentially intentional, I mean that in its classic sense: how we understand computing is permeated by assumptions of non-effective intentional directedness towards potentially distal subject matters (numbers, professors, weather in Oaxaca, etc.). This means that the vocabulary shift has had the unfortunate consequence of depriving computer science of any way to refer to concepts and phenomena constitutive of its subject matter, and essential to any adequate theory of computation.

Second, the classical meanings continue to undergird public discourse, holding up society's understanding of computing and artificial intelligence. That is, the "pun" effected by the shift between the classical and modern computational meanings props us an ongoing public myth about how to understand

computational systems. When it is publicly stated that it has been proved that an employee recruiting system is not gender biased, or that the control program for a nuclear reactor is correct, the public will assume that this means not gender biased and correct *in society*, whereas in fact what has been proved will be various properties of data structures and other computationally-internal structures.[56]

Disentangling these confusions, and laying out the intentional nature of computing, are some of the tasks of this book.[57] I say these things here only to emphasize: (i) that I take computation, in spite of any surface distractions, to be intrinsically intentional, in the classic sense; and (ii) that computer science's redefinitions of traditional semantical phenomena in no way implies that computer science has either avoided intentionality in its subject matter or successfully naturalized what is intentional. Rather, it has merely served to hide computation's manifest intentional character from theoretical view.

In spite of its protestations, that is, I believe computing is at its core an indissoluble dialectical admixture of meaning and mechanism.

D Radical Naturalism

The current theoretical situation puts us in a predicament. If we do not recognize computing as fundamentally intentional, then we cannot give an account of it as computing. If we do recognize it as intentional, then we need a theory of intentionality in terms of which to frame it. But as suggested above, we do not have such a theory.

Properly understood, I argue in chapter 2, theories of logic *do* take logic to be intentional, albeit in the weak sense that they simply assume the existence of a mapping from syntactic expressions onto semantic values: onto objects, in the case of

---

[56] «…ref *Limits*…»

[57] They are also tasks I undertake in the context of reflection and self-reference in «…CR…».

terms; onto properties and relations, in the case of predicates and relation symbols; and for the syncategorematic categories of "purely logical" operators, onto conditions of negation, conjunction, quantification, etc. But it standard logical practice to take these semantic values as given—as parameters to the theoretical account, in what is known as "the interpretation function." No explanation of the origin or warrant for such interpretations is provided—no account of where they come from, or, crucially for our purposes, what it is that enables the sentences and terms to refer to entities and states of affairs far beyond their causal confines. The ubiquitous semantical interpretation function is understood to be essential to, but in no way explained by, the logical analysis.

Most philosophical theories of language and mind also recognize the intentionality of their subject matters, but their most developed versions are far too specific to the particularities of specifically human language and thought to be applicable to computation in general. And they too frequently fall under causalism's sway, accounting for neither the nature nor the warrant of their referential relation to what is distal and beyond effective causal reach—relations that are nevertheless determinative of meaning and truth. Even causal theories of reference, as I have suggested, fail to make the grade. They may be suggestive of one way that terms may acquire particular reference, but they do not explicate how the apparatus of reference arises in the first place, what holds it up, how agents have a sense of a world in which to place objects the objects that new terms are referring to ("dubbing"), how the world "makes a claim" on our thoughts and imagination and reasoning, and so on. Inferential theories of reasoning similarly tend to duck the issue of how reasons and explanations are about things in the world outside them—relations essential, perhaps among other things, to determining their claims on reason.

In sum, we lack a sufficiently encompassing and adequately naturalistic account of intentional systems in terms of which to formulate an adequate theory of computing.

If that were not enough, an even larger issue is at stake. Consider again Hall's[58] tri-partite classification of philosophical accounts of computation: (i) *mechanical* or purely syntactic accounts, which take computation to be a particular kind of physical system, independent of all intentional issues; (ii) abstract or *mapping* accounts, in which computation is view to be a property of an abstract (perhaps constitutively mathematical) realm, with concrete computers merely being physical devices onto which the states of computation are mapped, or by which they are realized or implemented (Dijkstra's view, shared by many computational theorists of mathematical bent); and (iii) *semantical* accounts, in which computing is directly characterized in essential part with reference to its ineliminably intentional character.

Though fundamentally different on what property or properties they to be distinctive and defining of computation, these three categories of account agree on one essential fact: they take computation to be **special** in some way—more restrictive than the general case. Adherents of the first view do not take "being computational" to be a necessary[59] property of all mechanical systems; advocates of the second do not assume that all mathematical structures are computational; and proponents of the third do not view all semantically evaluable systems as computers. Each view is committed to the idea that not everything in the realm in which they locate computing is computational, in other words; they all take computational systems to be a proper subset of the realm under discussion. Pan-computationalism, that is, to put it bluntly, is in each case viewed as false—and a theory that leads to that conclusion to be fatally flawed.

It is not hard to see why this specialness is viewed as so important. Only if computation is special does the computational theory of mind (CTM) have any substance, for example. Since

[58]«…Hall forthcoming…»

[59]«…logically necessary in the sense that everything would by necessity be computational, that is; not merely that, as a matter of empirical fact, it might be that all existing systems were computational…»

minds are manifestly intentional, if all intentional systems were computational, the CTM would necessarily be true—and, just as surely, vapid. Moreover, if 'computational' is not a theoretically substantive predicate, the question immediately arises as to what computer science's considerable and impressive body of results is actually about.

I believe that all these objections are mistaken or misconstrued. I give the briefest stab of some of the reasons in *Smith (19…*FOC*)*. But the bottom line is simple enough: I do not believe, in the final analysis—at the fundamental level we are interested in here—that the realm of computing a theoretically interesting subset of the full class of meaningful mechanisms (not, as philosophers might put it, a "natural kind"). Rather, I believe computing is a **site, not a subject matter**: an area where we are exploring the nature of the meaning and mechanism admixture—and doing our best to construct artificial instances of any configuration that we understand well enough to render in material form. Admittedly, what we have developed to date occupies a very limited and proscribed set of circumstances, but it still early days. For example, I am not saying that we have yet constructed systems that exhibit what has been called authentic intentionality[60]—that shoulder the full burden, even in context, of establishing the non-effective semantic relations to their subject matters. But on the day that we do, no vocabulary police will swoop in and warn us not to call the resulting systems computers. Computers are meaningful mechanisms, the best we know how to construct. Period.

Theoretically, this "site not subject matter" claim make the study of computing more interesting and more consequential, not less. It is a site where we (do, or at least ought to) articulate a theory of meaning and mechanism, and their dialectical interplay in concrete, occurrent systems.

Since I do not believe that an adequate account of meaningful mechanisms can be framed in causalist terms, this makes the

[60]«…refs…»

study of computing the site of an epic battle over the adequacy of the causalist construal of nature. Our current computational systems are radically more complex than the inclined planes and planetary orbits and molecular structures that unleashed the contemporary "natural" sciences. But just as surely they are radically simpler than full-blooded human intentionality. And perhaps, therefore, the computational site can be one whose intermediate levels of complexity can help us get a theoretical handle on the nature of the profound dialectical interplay.

To date, it hardly need be said, the battle has been engaged in almost wholly causalist terms. The "meaning" side of the dialectic is losing, epistemically. But perhaps we can change that.

Doing justice to computation, in sum, is a challenge to our current conception of science, at the very deepest levels. It is a challenge to the adequacy of causalist accounts of the world. And it is a challenge as to how to understand how the full realm of thought, meaning, language, reference—how to full understand how the gamut of intentionality fits into a palatable overall metaphysical picture of the world.

And not just into a metaphysical picture of the world, but into a naturalistic one. We do not have to abandon commitment to the double standard adumbrated at the outset: eschew everything spooky, anything that does not accord with our direct experience of the world; and reject reliance on any external authority, of human or divine inspiration, instead granting authority to (and only to) the world itself as the arbiter of truth, the grounds of reference, and that which science (if we continue to use that term) must hold sacred and to which it must defer.

I call this project **radical naturalism**, to distinguish it from the causalism that currently has purchase on the unadorned term. I take it as a condition of success on radical naturalism that it make room for, and allow explanation of, the sorts of non-effective relation constitutive of referential directedness, "aboutness," and our intentional life. We can use computation as an exploratory site for this study—a site in which to a develop an account of the inexorable interplay of meaning and

mechanism, one that recognizes the non-effective along with effective, and gives us an opportunity to draw out morals applicable to significant phenomena more generally.

This is what makes computation such an important object of study, worth a lifetime of effort. Perhaps by taking computation's measure as a site where physical mechanisms participate in the realm of meaning we can take a few small steps towards a palatable naturalistic reënchantment.