GOFAI Considered Harmful (And Mythical) [1]

Drew McDermott

Draft! Comments welcome! Do not quote!

2015-06-27

Abstract: The acronym "GOFAI," coined by John Haugeland to describe the research strategy supposedly embraced by early AI, has had staying power. I introduce the term *symbolicism* for this "paradigm" (which is closely related to Newell and Simon's *physical symbol system* hypothesis), although the theory, or ideology, behind it is much less rigorous than it first appears. Computers are not "interpreted formal systems" in any interesting sense, and AI programs, even theorem provers, do not rely heavily on formal systems. Hence the *only* thread that supposedly binds early AI programs is their use of semantically interpreted symbols. Unfortunately, Haugeland's treatment of semantics is defective, and never addresses how symbols are tied to the things they denote, because the concept of "interpretation" is mixed up with concepts that should be treated as purely syntactic. All this confusion, combined with the tremendous influence of Haugeland's work and that of Newell and Simon, has been an important factor in the turn away from representation and computationalism generally on the part of cognitive scientists. I close with brief sketches of alternative treatments of "robosemantics" and the history of AI.

Keywords: AI, GOFAI, history, semantics, representation, cognitive science

1 Introduction

One of the most enduring acronyms in cognitive science was coined by John Haugeland in his book *Artificial Intelligence: The Very Idea* (Haugeland, 1985).[2]

> . . . [N]ot just any intelligent artifact would be Artificial Intelligence—not in our sense, anyway. This is not to deny, of course, that there are many theories of the mind, including interesting non-AI mechanisms and computer models. The point, rather, is to maintain conceptual clarity by keeping tight reins

[1] With apologies to Edsger Dijkstra (1968).

[2] In citations, AIVI. Except where noted, any emphasis is always present in the original.

> on terminology. To mark the intended careful usage, I have capitalized "Artificial Intelligence" throughout; but, lest that not be enough, or if someone wants these words for another role, we may also speak more explicitly of what I shall call Good Old Fashioned Artificial Intelligence—GOFAI, for short. (AIVI, p. 112)

Haugeland identifies these as "the claims essential to all GOFAI theories" (p. 113):

> 1. our ability to deal with things intelligently is due to our capacity to think about them reasonably (including subconscious thinking); and
> 2. our capacity to think about things reasonably amounts to a faculty for internal "automatic" symbol manipulation.

Right away it is clear that this is really *one* claim imputed to the proponents of GOFAI, because the phrase "think reasonably" in clause 1 is immediately cashed out as use of the "faculty for internal 'automatic' symbol manipulation." We can paraphrase that claim as

> [O]ur ability to deal with things intelligently is due to our ... faculty for internal "automatic" symbol manipulation.

I will argue that what Haugeland means by "internal 'automatic' symbol manipulation" is so unclear that it is impossible to see how it might explain "our ability to deal with things intelligently." His crucial discussion of the meanings of symbols is wide of the mark; it fails to even try to explain how symbols can be *about* things. As a result of these flaws, his book, though widely cited, is poorly understood. It has been used by many different people to justify many different views of artificial intelligence. On the whole, it has exerted a baleful influence on the ties between AI and the rest of cognitive science.

Here's how the paper is organized: In section 2 I describe Haugeland's account of GOFAI in greater detail, explaining where I think it is mistaken either philosophically or historically. A subsection deals with the work of Newell and Simon on their "physical symbol system" hypothesis (Newell and Simon, 1976, Newell, 1980), which is often linked with Haugeland. In section 3 I discuss the remarkable influence Haugeland's and Newell and

Simon's ideas have had, even though everyone seems to have a somewhat different idea of what they meant, and conclude that the net effect of this confusion has been to "poison the well" for computationalism in cognitive science. In section 4 I propose a better way to think about formal systems and their semantics in the context of cognitive science. Section 5 sketches an alternative history of AI, with a different view of what's has and hasn't changed since the early period. Section 6 is a summary of the important conclusions.

1.1 Previous Work

In the five years after its publication, *Artificial Intelligence: The Very Idea* was reviewed several times, almost always favorably (Andrew, 1987, Ford, 1987, Ogilvie, 1986, Pohl, 1988, Richmond, 1991, Vellino, 1986, Eckardt, 1988). I have been able to find only two reviews that raised concerns. Neander (1988) praises the book, but is especially bothered by its neglect of analog computations, and even of some ordinary digital algorithms, such as those involved in computer vision.

> Yet as Haugeland ... notes, ordinary digital computers can and do manipulate images: rotating them, or comparing two of them to extract a depth analysis of the visual field, for example. Haugeland maintains that when they do this the process is not pseudo-linguistic and so it is not computation ([AIVI] p. 228). This is an ad hoc ruling, and one from which it follows that not all computing is computation. (Neander, 1988, p. 270)

I take a similar position here (see section 2.1).

The only really negative review of AIVI that I have been able to dig up is that of Flanagan (1986). His main target is Haugeland's attempt to explain how a computing system can succeed in referring to something, the problem of *original meaning,* or *naturalizing semantics* (Dretske, 1995), or *psychosemantics* (Fodor, 1988). He summarizes by concluding that "I don't see that he [Haugeland] has made a dent" in the problem.

I agree completely with this criticism, too, although I diagnose Haugeland's failure rather differently (section 2.2) and of course I am more hopeful about the eventual success of computationalism in naturalizing meaning.

All of these book reviews were quite short. A detailed critical treatment of AIVI is long overdue.[3]

[3]Because it took so long for this paper to gestate, John Haugeland died before I had written much down (except for a few cryptic remarks in McDermott (2001b, ch. 2)). So I

Another theme in this paper is the unproductive estrangement between AI and other areas of cognitive science, especially cognitive psychology. For an excellent essay going deeper into this topic, read (Forbus, 2010).

2 Haugeland's GOFAI

John Haugeland had a gift for a turn of phrase. The title of his book, *Artificial Intelligence: The Very Idea,* is a good example, as is, of course, "Good Old-Fashioned AI," not just because it makes a goofy meme, but because it connotes something that was once fashionable and is now remembered only by old people swapping stories on the porch.[4]

The acronym is not introduced until page 112. By that point, he has described computers in great detail, and a theory of what their symbol manipulations mean. A picture of a research paradigm is meant to have jelled. The discussion occupies two parts: Chapter 2, which describes *formal systems*, and the first half of chapter 3, which concerns the semantics of such systems. We start with the first of these, leaving chapter 3 until section 2.2.

Fans of AIVI will notice that there are whole chapters that will receive almost no attention in what follows. These are well worth reading, especially the description of Babbage's analytical engine in chapter 4, some details about early heuristic programs in chapter 5, and some useful discussion of possible limitations of computationalism in chapter 6. Some parts should come with a caveat, especially the section on the "frame problem" in chapter 5. This phrase is the label philosophers use to cover their diagnoses of brittleness in AI systems. A system is brittle if it breaks down outside of a narrow comfort zone. It's hard to diagnose the cause of a program's brittleness, or any other bug, when you don't really understand how it works. Philosophers' guesses are usually of little value. See McDermott (1987).

But most of this later material has little bearing on my focus in this paper, which is what, if anything, GOFAI is supposed to refer to.

---

lost my chance to get his feedback. When we were both young I always found John to be generous and supportive. He gave me the opportunity to read and comment on a draft of *Artificial Intelligence: The Very Idea*, and he graciously acknowledged my (tiny) input in the published version. I recall thinking the book was a mildly critical but possibly useful piece of publicity for AI. I heartily wish I had kept the conversation with John going, so that each of us could have been more sensitive to the ways the other's views were evolving.

[4] Me, for example.

2.1 Formal Systems

On p. 48 of AIVI, Haugeland defines a *computer* as an "*interpreted automatic formal system.*" In order to explain the meaning of this key phrase, he starts with a chapter on automatic formal systems (chapter 2), and succeeds pretty well in explaining what that part means (saving the word "interpreted" for chapter 3). The discussion has some faults, however. It emphasizes games, explaining lucidly the difference between digital games like chess and nondigital games like billiards. But in a nontrivial game or puzzle there is a choice of moves. A computer, however, is, or is described by, a *deterministic* formal system.[5]

In a nondeterministic system, there are many sequences of "moves" that can be made. The resulting states are all "valid" in some sense, but typically only a minority, often a tiny minority, are useful for governing behavior. It is easy to create a formal system to generate all possible continuations from a chess position, but most moves will lose to a good opponent. It is almost as easy to generate all proofs in formal number theory, but most of the proofs are of uninteresting theorems. A computer simulates such a nondeterministic formal system by exploring some of the paths down which the nondeterministic system might go and ignoring the rest. It backs out of blind alleys and stops when it finds the best answer, or a good-enough answer.[6] This process goes by the label *search*. Sometimes (as on p. 76), Haugeland speaks of the legal-move generator as a *player,* using the term *referee* for the part of the program that handles the search (and decides when the game is over, among other tasks). But he seems to want these categories to apply to all computers, as we will see shortly.

Throughout AIVI, Haugeland seldom distinguishes between *computers* and computer *programs*. It is true that the distinction is often not terribly important; a program can establish a *virtual machine* that we can pretend is the actual formal system of interest. But in this context the ascent to virtuality is used to duck the issue of how a physical computer can *instantiate* a formal system. Not until chapter 4 does the book get down to describing real computers. (Chapter 5, entitled "Real Machines," is actually about real *programs*; the same maneuver again, with an even more obscure purpose.

In the discussion of formal systems in chapter 2 it is always assumed that

[5]The truth is more complicated than was apparent in the 1980s. Modern computers often consist of multiple cores without a well defined global state, and hence no useful description as a state-transition system (Patterson and Hennessy, 2009). But for expository purposes we can accept this simplification; the full truth does not alter the point being made.

[6]See any intro-AI textbook, such as (Russell and Norvig, 2010, ch. 5) for the details.

some agent is producing and manipulating *tokens* from an *alphabet*. The agent must use finite resources to do these manipulations, and they must not require intelligence or insight. All manipulations must be reducible to a finite number of primitive operations. So all the agent has to be able to do is to carry out those primitives, and to combine them into sequences, conditionals, and loops. (The agent also requires a storage medium, possibly an unbounded amount. Turing (1936) gets down to the nub of the issue much quicker, but of course assuming a higher degree of mathematical sophistication in his audience.)

To create a digital computer, to make a formal system *automatic*, in other words, Haugeland has to get the anthropomorphic agent out of there, and replace it by an electronic circuit. He never actually does. Instead the player and referee stick around, the players being components that perform arithmetic operations, such as multipliers and adders, and the referee being the part that finds the next instruction and decodes it. "What do algorithms have to do with automatic formal systems? In a way, everything. ... Suppose each primitive operation gets assigned a single player ..., and suppose the referee has all the primitive recipe-following abilities. Then, roughly, the referee can follow the recipe, signaling the players in the right order ..." (AIVI, p. 80).

On the same page, he says, "But what about these primitive abilities? Well, whenever an ability can itself be specified by an algorithm (in terms of simpler abilities), then that whole ability can count as a primitive in a higher-level algorithm. The only restriction is that all the required abilities eventually get analyzed into ground-level abilities that are *really* primitive—abilities so 'mindless and mechanical' that even a machine could have them" (AIVI, p. 80). "Connection to physics is made at ... the level of operations so primitive that they can be carried out by mindless mechanisms" (p. 82). But so far the treatment of the machine has been at the level of metaphor; it is imagined as a person playing or refereeing a game. The neophyte reader is not sure that there is such a thing as a "mindless mechanism," or in what it might consist.

The discussion has perhaps been useful for readers interested in basic algorithms, for tasks like multiplication of long numbers. But it has left us confused about what role nondeterminism plays in the operation of a computer or program, and what fills the role of "referee" in a real computer. Many readers must conclude that all programs, or a large fraction, are devoted to performing search through nondeterministic alternatives.

But computers do lots of other things, obviously. Suppose we connect a computer to motors controlling a ping-pong paddle — a real, physical one.

The computer also has a camera giving it a view of the ping-pong table and the ball in flight. The signals coming from the camera are digital, and the signals sent to the motors are digital.[7] It plays ping-pong fairly well. Is the program a formal system? If not, why not? Ping-pong doesn't fit Haugeland's (or anyone else's) notion of a formal or digital game, but here we have a digital system playing it. What went wrong?

The problem is that the metaphor of a formal game got used in too many ways. Sometimes it refers to a *problem domain* a program is trying to solve. Here it matters that the domain be characterizable as a nondeterministic formal system. But sometimes it just refers to a *program,* such as one for multiplying Arabic numerals, or one for choosing a good chess move — or tracking a ping-pong ball through a series of images. If Haugeland wants to say that there is always a "referee" and a "player" in a digital system, those terms are completely divorced from the game being played; they are just similes for the instruction decoder and branch logic and the arithmetic operations, and not very good similes. The "referee" in a deterministic computer does not make choices or enforce rules; it decodes instructions and comparator states. Even here, the word "decodes" connotes too much. The bits of an instruction code are run through the electronic gates of the decoder and out the other side come bits that allow information to flow, or block information from flowing, through other circuits. There is no longer any question of a formal game being played. The program in question might play an informal game like ping-pong, or do some other arbitrary task (cf. Neander (1988)).

Just before the acronym GOFAI is introduced, Haugeland tries to straighten this confusion out and only deepens it. (We're skipping ahead to chapter 3, briefly.) He draws a distinction between Type A and Type B systems (AIVI, pp. 110–112). Type As are those for which there is a characterization of the "legal moves" by way of a formal system, plus a referee that searches among the legal moves. This module is characterized with these phrases (all from AIVI, p. 111):

- "automated in a sophisticated manner"
- (selects legal moves that are) "not ... merely ... intelligible, but also ... clever, well motivated, interesting, and all that"
- "chooses something interesting or clever" (from the legal moves)
- "superimposed level of wit and insight"

[7] Reality is continuous, so the boundaries around the digital realm are crossed by transducers and effectors that transform energy into digital signals or vice versa.

A Type B program is one for which no formal system can be extracted. The only description we have left are those vague phrases, and this: "The really exciting prospect ... is building systems ... that are not restricted to domains where a suitable division (formalization) can be found" (AIVI, p. 111). It seems that so far we have established absolutely no constraints on intelligent programs except that they *be* intelligent. And Haugeland acknowledges that *any non-GOFAI system can be simulated using a computer.* "... Should any ... non-GOFAI ... theory of intelligence ... succeed, then it ought to be possible in principle to build an intelligent artifact that works as the theory says.... It should also be possible ... to *simulate* that intelligence on a computer" (AIVI, p. 112). But simulating a system does not mean instantiating it, *unless it manipulates symbols.* "The thesis of GOFAI, however, is not that the processes underlying intelligence can be described symbolically (a feature shared by hurricanes and traffic jams), but that they *are* symbolic (which differentiates them utterly from hurricanes and traffic jams)" (p. 113).

To characterize his argument somewhat fancifully, Haugeland starts by cranking up the magnification on his microscope so that crystalline formal systems occupy almost all of the field of view. He then zooms slowly out until the formal crystals are revealed to be a few grains of sand on a pair of flip-flops after a trip to the beach; and finally, when describing Type B systems, reaches in and brushes the sand out of the picture completely. And yet the idea that he's talking about a vast and austere Sahara of formal systems hangs in the air, when all he's really left with is a pair of flip-flops. The concept of symbol has drifted free from the formal context where it started, yet many of the connotations of formal systems are still supposed to apply to symbol-manipulating programs.

Curiously, according to the index, there is absolutely no mention of symbols back in Chapter 2. Yet all of this material concerns manipulation of 'tokens," and in the description of what it means to be digital, this idea is hammered on. "A *digital system* is a set of positive and reliable techniques (methods, devices) for producing and reidentifying tokens, or configurations of tokens, from some prespecified set of types" (AIVI, p. 53). These tokens are not symbols, however. How could they be? If digital systems intrinsically produce symbols, then a robotic ping-pong player would spew them forth in great abundance.

The concept of symbol bears a heavy weight for something so obscure.

2.2 Symbols and Semantics

Let us put the syntactic confusion about symbols aside, because there is another source of great puzzlement we need to deal with. Remember that a computer is supposed to be an *interpreted* formal system (p. 48). Chapter 3 of AIVI is aimed at explaining this part of the definition.

Haugeland begins by entertaining the idea that thoughts might be like linguistic utterances. Immediately the question arises what the thoughts *mean*, how they are to be interpreted. Haugeland argues that "the basic principle of all interpretation is *coherence*.... There are two sides to the coherence principle, one for what you start with (your 'data') and one for what you get (your 'renditions' of the data). Specifically, you must start with an ordered text, and you must render it so that it makes reasonable sense" (AIVI, p. 94).

I think this is quite right, but it overlooks an important step: finding the "texts" in the first place. What you start with is a set of physical situations which must be mapped onto the signals they are supposed to be. A classic example is the hieroglyphic writing system used by Egyptian priests for inscriptions on royal tombs and monuments. The people who understood the system died out, or rather their culture did, and for centuries thereafter hieroglyphics were viewed as a mysterious pictographic code that would reveal vast wisdom if successfully deciphered (Pope, 1999). But would-be decipherers usually misparsed the pictorial elements of the script. They didn't realize which groupings of elements constituted the characters of the alphabet. They didn't realize that it *was* an alphabet.[8]

Let's use the term *decoding* (McDermott, 2001b) for a way of mapping physical state types to abstract symbol strings.[9] The first proper decoding for hieroglyphics was worked out only in the nineteenth century, by Jean-François Champollion (Pope, 1999).

Or consider Haugeland's example of potato-shaped Plutonians whose "tiny leaf motions turn out to be ... translatable conversations, about us, the weather, astrophysics, and what have you" (AIVI, box 3, p. 96). For the leaf motions to be translatable, they first have to be recognized as a language of some kind. Some aspects of the motions are important in defining the symbol types composing that language, and some are not.

[8] Adorned with various nonalphabetic features.

[9] There's no reason to confine ourselves to linear sequences of characters, rather than arrays, trees, or some other arrangement, but for theoretical purposes strings will suffice. Or so Turing argued when he made the tapes of his machines one-dimensional (Turing, 1936). I believe a complete theory would have to include analogue signals as well; see section 4.

Haugeland's key example is of someone trying to decide among three different decipherings of a cryptogram (AIVI, table 1, p. 102). Only the third makes the cryptogram into a set of true mathematical statements. (The first makes them gibberish, and the second makes them into a set of *false* statements.) The coherence principle then states that the third deciphering is the correct one, because it makes the most "sense." (The null hypothesis, that a long random string of characters has a one-to-one mapping that produces a set of true equations, is ridiculously improbable.)

This example is chosen, I think, as an idealization, but it idealizes too severely. The alphabet used on the encoded side is the standard Roman uppercase letters. On the decoded side it is assumed to be decimal digits plus arithmetic symbols such as "×" and "=". Where this constraint came from is not clear.

If we care about what brain patterns or mental representations mean, then two elements are missing:

1. Nature does not tell us in advance what aspects of the physical situation form repeatable patterns; the relevant decoding must be inferred.

2. The cryptogram example stops when true arithmetic statements are found. In a realistic case, that would just enable us to ask the next question: What is being *counted* or *measured* by the numbers being added, subtracted, etc.? Is the person computing $8 \times 8 \times 8$ trying to guess the number of jellybeans in a cubical box?

The letter cypher is described as a "strange game" that Athol has "discovered" (AIVI, p. 100), but he has obviously not discovered it in the sense that one normally discovers games. He has apparently come across some enciphered material. When deciphered, the result is a list of simple arithmetical truths such as $8 \times 8 \times 8 = 24$ or $2/2 = 1$. It's unlikely that someone would write down "$2/2 = 1$" for any purpose except homework, so it seems that some bored child has enciphered their homework.

From this example the conclusion is what Haugeland calls the "Formalist's Motto":

> If you take care of the syntax, the semantics will take care of itself.

But there's a sense in which he hasn't even started to discuss semantics. I'll demonstrate by examining his other primary example, a cryptogram

Abcd aefg ahf ijkkg

that can be deciphered, assuming a letter-permutation cipher, in only three reasonable ways (AIVI, pp. 95–97):

> Dumb dogs dig wells.
> Cold cats cut riffs.
> Rife rats rot hulls.

(Two unreasonable decipherings are "Volt viny van ruddy" and "Zyxw zvut zsu rqppt.") His point is not that cryptograms are ambiguous, but the contrary, that the longer the cryptogram the less likely it is that there exists more than one deciphering that looks like a message someone might want to send. In Haugeland's short and contrived case even the syntactically correct decipherings don't look much better than "Volt viny van ruddy," but the probability of recovering a syntactically legal English sentence by chance is so low that if you find a decoding that pulls the trick off, it's probably right. (If you can't think of any reason why your adversary would talk about dumb dogs or riff cutting, that's okay; there's probably another layer of code to get through.)

Analyses such as these do not begin to penetrate the real problem of semantics, which is to explain how symbols can mean something even if there is no audience (Fodor, 1988). How does the principle of coherence apply to symbol systems in the heads of animals or robots? If brains are what animals use to compute things, presumably the symbols they use refer to things in their environment, such as terrain features, familiar places, and other animals.[10] Fodor (1988, p. xi) puts the problem thus: "How can anything manage to be *about* anything; and why is it that only thoughts and symbols succeed?"

The distinction I'm making will be clarified by a more realistic example of cryptanalysis, drawn from the story of the Allies' success in breaking German ciphers in World War II (Hinsley and Stripp, 1993). For reasons I need not go into, a crucial technique in figuring out the fresh settings of German ciphers every day was to look for expected pieces of plaintext ("cribs"). One of the weirdest techniques used was to send out airplanes to lay mines in places where the Germans were likely to find and clear them. The Germans were then sure to transmit reports of fresh mines being seen in these areas. If the British used as cribs the word *Sperrnachrichten* ("minefield alert") and the coordinates where the mines had been laid, they could infer the cipher settings.[11]

[10] And object properties, animal species, and other abstractions, but I am neglecting this dimension of the meaning problem, and several others.

[11] This was called Operation Garden by the cryptanalysts (Morris, 1993, p. 235).

The point is that semantics started after a message had been deciphered to yield a series of German words. Behind the words lay the entities in the world that the words referred to, particular mines and particular geographical locations. Of course, the British cryptographers didn't care about mines they themselves had laid; they wanted to know about the locations of German wolfpacks and *Fliegerkorps*.

Fodor's conundrum now presents itself. The meanings of wartime ciphers, or any other ciphers, are ultimately set by their users. A pattern of signals refers to a set of coordinates because the transmitter intends them to, and the receiver expects the transmitter to have had that intention; they are working from the same protocol. But inside an autonomous computational system there is no one corresponding to the sender or transmitter. One might argue that the programmer is the ultimate arbiter of what the symbols used by programs mean, but that solution will not work if we want to treat animals as autonomous computational systems.

One conclusion is that therefore animals (especially the human kind) are not computational systems. This conclusion is often drawn in more or less this way (e.g., by Searle (1990), McGinn (1991), Jacquette (2009)), as I'll discuss in sections 3 and 6. The opposition has had trouble making itself heard. There are two main reasons for this.

1. Under the influence of Haugeland's analysis, and (unfortunately) a faulty but influential analysis of cognitive psychology from within the field by Allen Newell and Herbert Simon (1976), the excitement caused by neural networks in the 1980s (Rumelhart et al., 1986) led to the idea that the computations performed by brains belong to a different paradigm from the "physical symbol systems" Newell and Simon thought were important. (See below, section 2.3.)

2. The philosophers most concerned with "psychosemantics," such as Fodor (1988, 1990) and Dretske (1981, 1995), muddied the waters by mixing up the theory of symbol meanings with the theory of word meanings.

I'll spend most of my time discussing the first issue, and just say a bit about the second problem here. It may seem harmless to mix up the theory of mental symbols with the theory of words. Aren't words the central exemplar of mental symbols? No. For one thing, there are probably many more internal symbols than there are words. Although many symbols have no meanings in the sense we are interested in — like parentheses, they do not refer to or mean anything — the number of meaningful internal symbols

must still dwarf the number of words (Millikan, 1984). Second, the focus on words tends to foreclose the possibility of nonhuman animals having symbols. Third, anything we can become conscious of, including words and their occurrences, can be distorted in the peculiar way we induce by filtering some beliefs through self-models (Graziano and Kastner, 2011, Metzinger, 2003, McDermott, 2001b). We think we *know* the meanings of words. To put it in the first-person singular, I seem to be the arbiter of what my words mean to me. Whether this is true or not for words (and I doubt it is), it is certainly not the case for symbols whose existence I am not even aware of. I think I know the meaning of "cow," which leads me to believe that I know the meaning of COW (the internal symbol tokened when I have cow perceptions or think cow thoughts). My certainty that "cow" doesn't mean "cow or horse seen in fog" (the standard example of the *disjunction problem* (Fodor, 1988)) is irrelevant when it comes to figuring out what a given internal symbol means. In other words, I deny Cummins's dictum that a theory of meaning "... must provide an account of what it is for a system to understand its own representations: Intentional states of S = contentful states of S that S understands" (Cummins, 1983, p. 68).

2.3 Physical Symbol Systems

The idea that AI depends on symbol manipulation is usually credited to Allen Newell and Herbert Simon. In their Turing Award lecture of 1975 (Newell and Simon, 1976), "Computer Science as Empirical Inquiry: Symbols and Search",[12] they named and conjectured the truth of the *physical symbol system hypothesis,* that "a physical symbol system has the necessary and sufficient means for general intelligent action" where a "physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure)" (CSEI, p. 116).

> A physical symbol system is a machine that produces through time an evolving collection of symbol structures. Such a system exists in a world of objects wider than just these symbolic expressions themselves.
>
> Two notions are central to this structure of expressions, symbols, and objects: designation and interpretation.

[12]But of course they distilled this lecture from two decades of work. In citations, I'll refer to this paper as "CSEI."

> *Designation.* An expression designates an object if, given the expression, the system can either affect the object itself or behave in ways dependent on the object.
>
> In either case, access to the object via the expression has been obtained, which is the essence of designation.
>
> *Interpretation.* The system can interpret an expression if the expression designates a process and if, given the expression, the system can carry out the process.
>
> Interpretation implies a special form of dependent action: given an expression the system can perform the indicated process, which is to say, it can evoke and execute its own processes from expressions that designate them. (CSEI, p. 116)

The meaning of this passage is terribly obscure. In truth, Newell and Simon did not have anything to say about how a physical symbol system (PSS) could refer to the "wider world" of objects outside the system, but they did believe they had achieved *some* understanding of some of the issues involved in semantics, as I'll explain. But choosing the words "designation" and "interpretation" seems perverse given that these terms have been used by formal semanticists in ways quite different from what Newell and Simon are talking about.

It's clear how *within* the computer one expression can refer to another (allowing the system to "affect the object itself"). The simplest such relation is that between a short (say, 32-bit) binary string and the location addressed by that string, a relation defined by the circuitry of any modern computer. A bit string used this way is called a *pointer*. Because the location addressed may be the start of a block of data, including more pointers, such a binary string can stand for a block of data of arbitrary size and flexible boundaries. There isn't much more to the idea of *data structure* than this. I don't think Newell and Simon would disagree that we can treat their words "symbol" and "expression" as synonymous with "pointer" and "data structure," respectively.

These simple, interlinked concepts make it possible to write a universal algorithm that interprets an appropriate data structure as a program. The data structure then denotes a procedure, as Newell and Simon suggest. "A physical symbol system is an instance of a universal machine" (Newell and

Simon, 1976, p. 117). In (Newell, 1980), Newell's amplification of CSEI, he is even more specific: the programming language "... Lisp is a close approximation to a pure symbol system..." (p. 138). Remember that Newell and Simon are arguing that for a system to be intelligent it is *necessary* that it be a PSS. Their subject matter was conscious problem solving (Newell and Simon, 1972), and one can argue that, given enough instruction (plus pencil, paper, and literacy), a person can carry out an arbitrary procedure.[13] But it seems clear now that this was a mistake if it was intended to provide a foundation for psychology. It would be more accurate to think of the brain as comprising many neuron populations with specific computational behaviors and the ability to learn, within narrow ranges (Botterill and Carruthers, 1999).

If the brain is neither digital nor programmable in any important sense, why are algorithms so useful in expressing cognitive-scientific hypotheses? There are a few possible answers, any of which can be correct, depending on context:

1. Algorithms can be solutions to AI problems that are divorced from any connection to psychology.

2. If one accepts David Marr's 1982 characterization of computational methodology in psychology, there is a level of algorithmic analysis between domain analysis and implementation.

3. Algorithms can be used to express parallel (multi-thread) algorithms as well as serial ones. The brain might contain a system consisting of several (or perhaps very many) subsystems running in parallel, and an algorithmic notation can capture how they interact, at some level of abstraction.

4. More loosely, an investigator might conjecture that a computer, consisting of one or a few elements running very fast, might perform a huge number of operations mostly serially, while a brain, consisting of a "massive" number of elements running slowly but in parallel, might perform the same operations (in some sense) simultaneously. The serial version of the algorithm might help prove that performing those operations would be sufficient for the task at hand, bolstering any direct evidence that neural assemblies in the brain actually do perform them.

[13] Or perhaps they had in mind plans that people figured out for themselves and then executed repeatedly; these may not, however, be a universal basis for computation.

It's important to realize that programs are not constrained to representing logical rules, nor do Newell and Simon ever claim that they are. It is curious that so many of their readers have come away thinking they did (see section 3). One reason may be that the first application they presented to the world was the LT program — "LT" for "Logic Theorist" — which proved theorems in the propositional subset of the formalism of *Principia Mathematica* (Newell et al., 1957, Newell and Shaw, 1957). It is easy to forget that the memories and CPUs of computers available in 1955 were so small and slow that there weren't that many domains to choose from, and propositional inference was one of the few.[14] Other than that, the domain had no special significance for them, and they tried to generalize as soon as they could to broader classes of problem (Newell et al., 1959).[15]

Newell and Simon also declare that a physical symbol system is a "*sufficient* means" for intelligence. I suppose they meant that once you had a PSS you could program it with any algorithm required for intelligence. This is a blunder, akin to saying that there is nothing more to materials science than the periodic table. I can't imagine what they were thinking, except possibly that the ability to program in a high-level language made things so much easier than programming in machine language that there wasn't much else to say.[16] If this is what they meant, it's belied by Newell's own research later in his career into even higher-level notations for procedures, such as production systems with general chunking mechanisms (Newell, 1994).

However, these missteps by Newell and Simon should not distract us from their contributions to computer science. They were the first to see the key role of data structures and search in getting computers to perform tasks that were considered "intellectual." They took for granted that symbols, pointers, and data structures could be found in the brain, a position that has become unfashionable (see section 3), but that is still in the running (Fodor, 2008, Gallistel and King, 2009). What they did not do was illuminate the

[14]LT ran on the JOHNNIAC computer at the RAND corporation. This computer had about 20 Kbytes of magnetic-core storage and another 46 Kbytes of magnetic-drum memory (Newell and Shaw, 1957), which held the program.

[15]By the 1970s, computers were bigger: most sites on the ARPAnet, which included all American AI labs, had PDP-10s with 2 Mbyte of main memory (Wikipedia, 2015). The average millenial carries a much more powerful computer in their pocket. It is odd how seldom it is suggested that the main reason we studied "microworlds" back then was that we only had micromachines.

[16]Compare the similar exuberant statement by the inventors of the FORTRAN programming language: "So far, those programs having errors undetected by the translator have been corrected with ease by examining the FORTRAN program and the data output..." Backus et al. (1957, p. 197).

semantics of symbols.

2.4 Why Interpreted?

We return to Haugeland's proposal that a digital computer is an *interpreted* formal system. One way to analyze this is simply to parse "interpreted" as "instantiated." As discussed in section 2.1, a computer synchronized by a single clock may be described by a formal system that describes its behavior in discretized time as a function of the inputs it receives. The computer, so long as it doesn't break down, *is* a model of the formal system in that there is a mapping from the formal system to the physical one that makes the formal system "true," in the sense that its initial state corresponds to an actual physical state and its rules are "truth preserving" — they never produce a state that does not follow from an actual state (Smith, 1988). I put quotes around the words "true" and "truth preserving" because the formal objects in question need not be of the type that we normally associate with those phrases.

Unfortunately, this sense of "interpreted formal system" does not seem to be the sense Haugeland cares about, or that anyone should care about. Normally the phrase refers to a formal system *plus an interpretation* chosen by a person. The claim that a computer is an interpreted formal system *all by itself* is intelligible only if the computer is a model of a formal system as just described, but, when a computer does a job for someone, the semantics of the symbols it manipulates have nothing to do with the semantics outlined above. If a computer signals a chess move in a game with a human opponent, what people care about is the semantics in which the signal denotes a move in the ongoing game, a move the machine (or, usually, its human handlers) are committed to. This level of symbol and semantics obviously has nothing to do with the level in which state-transition rules denote transformations of bits by circuitry and output of ASCII characters.

Curiously, in an earlier paper Haugeland (1978) had made exactly the same point in discussing Lucas's (1961) argument that Gödel's theorem makes AI impossible: "The error here is in neglecting that an interpretation of an object as a mathematician and an interpretation of it as a theorem-proving Turing machine would be on vastly different dimensions. The outputs put forward as proofs and truth claims on one bear no interesting relation to those put forward on the other" (Haugeland, 1978, p. 220f). This earlier paper by Haugeland explores many of the same concepts and claims as AIVI, only transposed to the domain of cognitive psychology, and addressed to a more sophisticated audience. It insists just as strongly that

semantics be an intrinsic property of a symbolic system: "A *quasilinguistic representation* is a token of a complete type from an intentionally interpreted articulated typology" (Haugeland, 1978, p. 218 (emphasis in original)). The densely phrased predicate here may be unpacked thus: An articulated typology is (roughly) a language; a token of a type from that typology is a physical instance of a string from that language. The requirement that the typology be "intentionally interpreted" *before its computational role and causal connections are delineated* corresponds to the requirement that a computer be an "interpreted automatic formal system" (AIVI, p. 48).

Following this clue, we can find another way to understand Haugeland's requirement by returning once more to the concept of a Type A program (AIVI, p. 110f) and assume he had in mind the classical member of this category: a theorem prover that operates by putting axioms together to make deductive conclusions.[17] If the axioms are about chess, then the conclusions might include statements about legal moves, or perhaps even good moves. Now suppose the theorem prover is connected via input and output transducers to the board, the input transducers producing new axioms about the opponent's current move, the output transducers making conclusions about the computer's next move true by actually moving pieces.[18] These transducers enforce the desired interpretation of the formal system.[19]

This chess-playing, theorem-proving robot is close to being an interpreted formal system in Haugeland's sense, but not that close. Even though it's a type-A program, the code that runs the transducers and makes the choices of which inferences to try is not itself part of the formal system, and is usually larger than the formal system. The system is not "interpreted" because it is a computer or because it is Type A, but because of its sensors and effectors.

We've come to the end of section 2, the central argument of this paper, so I'll summarize the critique of GOFAI:

1. GOFAI is supposed to be a paradigm for AI research. It is supposedly

[17]Indeed many theorem provers did and do work this way. Most of the exceptions are interactive theorem provers, in which a human is a partner in finding proofs of theorems.

[18]The classical AI system that comes closest to this picture is Shakey, the robot developed at SRI by Fikes et al. (1972). Shakey didn't play chess, but did solve problems involving pushing and stacking boxes.

[19]A simple chess-playing computer with no effectors and sensors except the keyboard still succeeds in referring to the pieces and making moves, I would argue, even though the operator does the actual seeing and manipulating. But explaining why exactly is beyond the scope of this paper.

based on the idea that human thought is ultimately analyzable as a set of formal systems, for which the computer, the ideal formal system, is the perfect simulation vehicle.

2. Upon inspection, the idea breaks apart into a bunch of concepts that don't fit together. On the one hand, digital computers are supposed to be formal systems *intrinsically* (AIVI, p. 48); on the other hand, even non-GOFAI systems can be simulated on a computer (AIVI, p. 112).

3. A large class of GOFAI programs are what Haugeland calls "Type B": they don't contain *any* identifiable formal system defining "legal moves." There are no constraints placed on a type-B algorithm, except that it must *manipulate symbols*.

4. This last-ditch constraint evaporates because the concept of symbol is so nebulous. On the one hand, digital computers seem to produce symbols by their very nature (AIVI, ch. 2). On the other hand, nothing is said about whether there are interesting thinking systems that don't compute, don't use symbols, or both. The prime candidate is the brain, of course, but nothing is said about whether the brain uses symbols, and if so what symbol tokens might look like in a brain, or whether a brain can do without symbols entirely.[20]

5. Haugeland leans heavily on the requirement that symbols and computations be imbued with meaning at a low level. But little is said about what would give a token a meaning. The Formalists' Motto ("Take care of the syntax and the semantics will take care of itself") is a straw man. A chess-playing robot must work to maintain the link between the symbols and the physical pieces; semantics never takes care of itself.

6. In the end all this is not surprising, given John Haugeland's philosophical orientation. He believed the solution to the problem of "original meaning" (see below, section 4) lay in the practices of human communities (Haugeland, 1990). Robots are left out in the cold.

[20] (Haugeland, 1978) is more forthcoming on this issue, discussing the possibility of hologram-like models accounting for some features of the way the brain seems to work.

3 The Influence of Haugeland and Newell and Simon

It would be difficult to overestimate the influence of Haugeland's 1978 and 1985 analyses, abetted by those of Newell and Simon (1976, Newell (1980)). These papers have had a steady stream of citations, continuing to this day.[21] For example, Nowachek (2014) describes the physical-symbol-system hypothesis and GOFAI as underwriting a theory of learning as "an operation of formal rule-based symbol manipulation" (p. 67). There is nothing unusual or noteworthy about this paper except that it's recent and not atypical. Those who find the GOFAI theme illuminating tend to be critical of those who supposedly espoused it. Quoting (Nowachek, 2014) again: "After only a few years of operation the GOFAI research program came to a grinding halt due to its own ontological limitations and the increasing philosophical criticism of it" (p. 68).

For want of anything better, I will use the word "symbolicism" to mean "good old-fashioned AI" as theoretically characterized by Haugeland and by Newell and Simon. I introduce this somewhat awkward[22] term because the belief that there is a fundamental research paradigm here is so widespread, even though there is little agreement on what that paradigm is. Dreyfus (Baumgartner and Payr, 1995, p. 71) sees early AI researchers reenacting the "rationalist" tradition in philosophy, opposed to the "empiricist" tradition. Winograd and Flores (1986) also describe GOFAI as "rationalist," but they lump empiricism in with rationalism. (Both they and Dreyfus see Heidegger's worldview as an important alternative to symbolicism, although whether it's possible to have a Heideggerian AI is still unclear (Dreyfus, 2007).)

I should mention that Haugeland studied with Dreyfus, and acknowledges his special influence in (Haugeland, 1978). So why not include Dreyfus's 1972 book *What Computers Can't Do* in my tiny list of key sources of the theory of symbolicism? The problem is that Dreyfus doesn't supply much of an analysis of symbolicism (and doesn't use that name, of course, since I just made it up). In (Dreyfus and Dreyfus, 1988, p. 34) (with his brother Stuart), he summarized the physical-symbol-system hypothesis thus:

> Newell and Simon hypothesised that the human brain and the

[21] Haugeland's meme is so pervasive that the phrase "good old-fashioned AI" is often used without attribution, as if it were a piece of folk tradition. (An example is Chemero (2013).)

[22] Alternatives such as "gofaianism" are even worse.

> digital computer, while totally different in structure and mechanism, had at a certain level of abstraction, a common functional description. At this level both the human brain and the appropriately programmed digital computer could be seen as two different instantiations of a single species of device — a device that generated intelligent behaviour by manipulating symbols by means of formal rules.

But of course what Newell and Simon actually said was that a PSS was "a machine that produces through time an evolving collection of symbol structures" (CSEI, p. 116). The distinction is of importance, as I've argued in section 2. Describing something as "a machine" puts no constraints on what it does except being computable; whereas "formal rules" are a quite special case. But where Haugeland makes a sustained attempt to minimize the difference, Dreyfus just assumes they're the same. In the later paper, he and Stuart Dreyfus draw a contrast between symbolicist and connectionist approaches to modeling intelligence. They praise the latter,[23] but mainly, it seems, because once a neural net has learned something, it is often hard to say what it has learned in any terms except improved performance (Dreyfus and Dreyfus, 1988, p. 46). But lots of computer programs have this property; the programs that implement neural nets are just one example.

Furthermore, Dreyfus's attraction to late Wittgenstein and the post-Husserl phenomenologists, especially Heidegger, as theorists of the human mind does not provide much inspiration to cognitive scientists, because these philosophers had no interest or faith in the future of neurophysiology or physiological psychology. It has therefore been difficult to create a Heideggerian AI or a Wittgensteinian neuroscience.[24]

Another use of the word "rationalist" to characterize symbolicism comes from Cummins and Schwarz (1987, p. 47f), who describe it as "... the idea that having a cognitive capacity is instantiating a function that relates propositional contents, i.e., a function that takes propositional contents as arguments and values and relates them as premises to conclusion. A cognitive system, in short, is an inference engine, a system that merits an inferential characterization. Thus, to explain cognition is to explain how a system can merit an inferential characterization, i.e., to explain how it can reason."[25] A symbolicist system's "... behavior is cogent, or warranted or

[23]

[24]For an attempt at the former see Agre (1997); at the latter, see Shanahan (2010).

[25]In the footnote immediately following this quote, this view of symbolicism is attributed to AIVI and a handful of other writers, although it's not clear which authors Cummins

rational relative to its inputs and internal states" (Cummins and Schwarz, 1987, p. 47).

As this quote shows, there is a remarkable muddle in the literature about the differences between *programs, algorithms*, *rule-based systems, inference engines*, and *axiomatic systems*. The last category is the most formally and precisely characterized, but also the most restrictive. As in my example of a theorem-prover-based robot, in practice such "Type A" systems almost always end up sliding into "Type B" systems, i.e., computer programs, the *least* restrictive classification.

Cummins and Schwarz founder on these rocks. "Orthodox AI's[26] proposal is that cognitive systems are computational systems: a cognitive system is described by a cognitive function because it computes representations whose contents are the values of the cognitive function, and computes these from representations of the function's arguments" (Cummins and Schwarz, 1987, p. 48). But being "computational" is no guarantee at all that a system will be "rational," let alone that it will "merit ... an inferential characterization."

Those with a more friendly attitude toward AI, and less prone to perceive any "grinding halts" in its history, tend to take a somewhat different view of symbolicism. Stuart Russell and Peter Norvig, in the leading AI textbook state that, "GOFAI is supposed to claim that all intelligent behavior can be captured by a system that reasons from a set of facts and rules describing the domain. It therefore corresponds to the simplest logic agent described [in this book]" (Russell and Norvig, 2010, p. 1024). They point out all the other kinds of agents that AI has studied, which overcome, or may someday overcome, the problems of the simple "logic agent." But their view of what GOFAI is exactly is as subjective as everyone else's.

There is a strange hunger for paradigms and paradigm shifts (Kuhn, 1962) in the history of AI, by practitioners within the field and outside observers. The field is really too young for this sort of thing, one would think. Nevertheless, it has been all too common since the 1980s for cognitive scientists to discern alternative paradigms that replaced GOFAI, or can be expected to replace it, or ought to replace it. The early favorite was connectionism, whose renaissance in the 1980s is documented in (Rumelhart et al., 1986). Connectionism uses "massively" parallel networks of simple elements of a few types instead of a single CPU. The simple elements are supposed to be neuron-like, and the resulting systems are supposed to shed

and Shwarz classify as *advocates* of symbolicism and which as *theorists* of it.

[26] Although Cummins and Schwarz do cite AIVI, they don't use the acronym GOFAI.

light on real neural systems. In practice this goal is often secondary or absent; and the networks are almost always simulated on a digital computer. One might think these two points would have something to do with each other, but of course a neurologically plausible system can be simulated on a digital machine, and a neurologically *im*plausible one can created using lots of physical elements operating in parallel.

Connectionism is still alive and well, but it couldn't magically solve all the problems mainstream AI researchers had discovered. The actual literature on connectionism is heavily mathematical and algorithmic — not that easily distinguishable from any other work in computer science (e.g., ch. 11 of Hastie et al., 2009).

Possibly because neural-net research turned out not to be a clearcut "paradigm shift" after all, an even more radical alternative emerged, in the form of *dynamicism* (van Gelder, 1998, Spivey, 2007, Shanahan, 2010): the idea that differential equations, phase spaces, and the other tools of dynamic systems theory were the appropriate ways to think about cognitive science.[27] The variables in these differential equations describe the activation levels of populations of neurons.

According to van Gelder (1998), dynamicism is opposed to the "computational hypothesis (CH) that cognitive agents are basically digital computers" (p. 615). Newell and Simon and Haugeland are cited, and everyone seems to harmonize with the "remarkable level of consensus" that exists regarding the CH (p. 617). In a footnote, van Gelder elaborates: "The version of this consensus now most widely accepted as definitive is probably that laid out in [AIVI]. The account of digital computers here is essentially just Haugeland's definition of computers as interpreted automatic formal systems ..." (p. 628).

The muddle has popped up again. First, note that AIVI does not claim that brains are digital, only that they can be analyzed as manipulating symbols:

> If Artificial Intelligence really has little to do with computer technology and much more to do with abstract principles of mental organization, then the distinctions among AI, psychology, and philosophy of mind seem to melt away. ... Thus a grand interdisciplinary marriage seems imminent; indeed, a number of enthusiasts have already taken the vows. For their new "unified"

[27] This is essentially what Bickhard and Terveen (1995) endorse, under the name "interactivism."

> field, they have coined the name *cognitive science*. If you believe the advertisements, Artificial Intelligence and psychology, as well as parts of philosophy, linguistics, and anthropology, are now just "subspecialties" within one coherent study of cognition, intelligence, and mind—that is, of symbol manipulation. (AIVI, p. 5)

The failure to grasp the distinction between programs and logical rules can lead to bizarre but understandable arguments such as that of Jacquette (2009)[28] [footnotes mine in what follows; emphasis in original]:

> If the mind is machine, then it should be possible to build a computer with whatever properties cognitive psychology attributes to the mind. This is the hope of *mentalistic artificial intelligence,* which seeks to create genuine intelligence in information processing machines. [p. 88][29]
>
> ...
>
> There are two kinds of artificial intelligence modeling, known as *rule-structured programming,* and *connectionist* or *parallel distributed processing.* Artificial intelligence has traditionally used the same rule-structured system as other kinds of computer programming to instruct a machine to execute commands when specified conditions are satisfied. Typical lines of code in a programming language issue commands of the sort: IF X = 0, THEN GO TO [5] [p. 93].[30]
>
> ... A modern digital computer, as we recall, is a universal Turing machine, capable of performing any and all computable functions, including those computed by connectionist networks. The following conclusion therefore seems inevitable. If a connectionist system considered only as such can duplicate any psychological phenomenon, then there is a copycat rule-structured mentalistic artificial intelligence program that can duplicate the same phenomenon. Despite the practical advantages of connectionist systems, parallel distributed processing cannot avoid the philosophical objections raised against mentalistic artificial intelligence in rule-structured programming. [p. 102]

[28] Who cites AIVI, but not Newell or Simon.

[29] This is essentially the same as Searle's notion of "strong AI" (Searle, 1980).

[30] This "rule" seems to be a statement in the BASIC programming language.

I'm sure it has occurred to more commentators than are willing to admit it that if formal systems are the essence of GOFAI, and *all* computers are "interpreted formal systems," then the distinction Haugeland was trying to make collapses, and connectionist systems are as GOFAI as any other program. Jacquette is to be thanked for unintentionally making the *reductio* so clear.

The tendency to see GOFAI in every mention of computation has caused some cognitive psychologists to avoid the whole confused mess of concepts. The advent of fMRI technology (Raichle, 1999) for near-real-time brain scanning has provided an alternative for them to rush to (Aue et al., 2009). A typical significant result is the discovery that moral decisions use different brain regions depending on which sort of moral dimension is in question (Greene et al., 2004). The neglected question is what is happening inside these regions. For now fMRI experimenters seem content to analyze the brain as a system of modules connected by "wires." A "wire" is sometimes taken as carrying messages in one direction, with feedback for learning purposes, or as a bidirectional channel. But what is going on inside those modules? And what is the format of the traffic in the channels in each direction? If we cannot express the answers in the language of the theories of computation and communication, we must usually be satisfied with superficial explanations in terms of "association," "experience," "images," and so forth.

Here is an example from Joseph Haidt and Craig Joseph (2004) in the context of theories about the neuroscience of moral judgments, although the details are unimportant:

> A useful set of terms for analyzing the ways in which such abilities get built into minds comes from recent research into the modularity of mental functioning. An evolved cognitive module is a processing system that was designed to handle problems or opportunities that presented themselves for many generations in the ancestral environment of a species. Modules are little bits of input-output programming, ways of enabling fast and automatic responses to specific environmental triggers. In this respect, modules behave very much like what cognitive psychologists call heuristics, shortcuts or rules of thumb that we often apply to get an approximate solution quickly (and usually intuitively) (Haidt and Joseph, 2004, p. 59f).

Jerry Fodor (1983) is cited for the original notion of modularity of mind, but he must be fuming at the depths to which it has fallen. One of his prime

examples was the hypothetical module that turns sound sequences into syntactic trees when we hear someone utter a sentence. The computations required to carry this out are "fast and automatic," probably "heuristic," but they are not "little bits of input-output programming." I use the word "computations" here because there simply is no other sort of model in our toolkit with the requisite power. We must wield the same tools to explain anaphora resolution, presupposition analysis, face recognition, social-rank judgment, "theory of mind" (Gopnik and Wellman, 1992, Goldman, 2012), geographical orientation, and many more operations plausibly handled by specialized neural circuits. So why have many cognitive psychologists regressed to a simplistic view of what these modules do? The complete story is probably complicated, but at least part of the blame must rest with the confused sense in the cognitive-science community that computationalism was somehow refuted when GOFAI "came to a grinding halt."

I don't want to give the impression that everyone in the cognitive-science community became hostile to computationalism after the first blush of romance had worn off. Many philosophers, representing a diverse range of positions about the nature of representation have remained comfortable with the idea that computation is either the key to the mind or a useful tool for understanding it, including Georges Rey (1997), Daniel Dennett (2006), and Jerry Fodor (2008), as have several cognitive psychologists with a philosophical bent, such as Zenon Pylyshyn (1980, 1984). Those doing research into analogy (Gentner et al., 2001, Gentner and Forbus, 2011, Itkonen, 2015) must confront the subject of complex representations head on. And, of course, post-Chomsky linguists can hardly give up the idea that thought involves computing with symbolic structures.[31]

Not every critic of computationalism sees GOFAI and its supposed demise as crucial events in the history of cognitive science. Stevan Harnad (1990) blames the "symbol-grounding problem" for the turn towards "nonsymbolic" systems. He defines a "symbol system" as one that manipulates physical tokens of some kind on the basis of "explicit rules."

> The entire system and all its parts – the atomic tokens, the composite tokens, the syntactic manipulations both actual and possible and the rules – are all ... "semantically interpretable": The syntax can be systematically assigned a meaning e.g., as standing for objects, as describing states of affairs.[32] (Harnad,

[31] Endorsement of computationalism should not be equated with endorsement of AI, as Fodor has made abundantly clear (Fodor, 1978, 1987).

[32] While I agree with Harnad's emphasis on semantics, his requirement that the "rules"

1990, p. 336)

Harnad credits several influences for his concept of symbol system, notably Fodor (1975), Haugeland (1978), Pylyshyn (1980), and Newell (1980). He makes the same mistake as Newell, Cummins, and others in requiring a symbol system to have all the structure of a programmable digital computer, but grants that there are human thought processes that are symbolic. He arguese that the Achilles heel of this otherwise successful research effort is that there is no explanation of the semantics within the theory:

> Many symbolists believe that cognition, being symbol-manipulation, is an autonomous functional module that need only be hooked up to peripheral devices in order to "see" the world of objects to which its symbols refer (or, rather, to which they can be systematically interpreted as referring). (Harnad, 1990, p. 338)

Harnad thinks this belief on the part of "symbolists" is overly simplistic, and that symbolism without symbols "grounded" in the right way can't be part of a realistic theory of human cognition. I confess that I don't see exactly the problems Harnad sees (Harnad, 2001, McDermott, 2001a), but I agree that semantics requires attention.

4 Robosemantics

Haugeland got semantics wrong, and his mistakes have unfortunately been copied by others. In this section I'll try to suggest ways to correct some of his errors. Solving all the riddles of semantics is, I am grateful to report, beyond the scope of this paper.

As I said in section 2.2, Haugeland's semantic analyses often stop short of the important issue, because they seem to mistake *decodings* for *interpretations*. A decoding is a mapping from a set of possible physical situations (its *domain*) to a set of abstract structural descriptions (its *range*). The standard case is classifying physical state tokens as occurrences of strings drawn from a language, but I think it is a mistake to assume that this is the only case. Here are some other sorts of ranges that decodings can target:

- Real numbers, or probability distributions on real numbers. (Any analogue computer must be analyzed this way.)

(program) get the *same* interpretation as the symbols manipulated is the fallacy that I discussed in section 2.4.

- Cartesian products of ranges. (E.g, when the physical state consists of frequency-multiplexed signals that are limited in bandwidth but otherwise independent.)

- Isolated discrete symbols *not* drawn from a language. Of course, we can always consider an isolated symbol to be a token of a length-1 string from a degenerate language; but we can just as well think of a more complex language as the result of product operations applied to isolated symbols.

In Chalmers's (1994) useful analysis of the concept of computation, the function he calls simply f is a decoding. His analysis is focused on the digital case, and for this case it seems exactly right to me. Once the concept of decoding is in place, it can be applied at many loci inside a computational system: inputs, outputs, storage elements, and subcomponents that are themselves computational. (There are all sorts of improvements we might pursue, such as incorporating state information into F, but they're not relevant here.)

Decodings are cheap, and worth every penny. A given physical state type can be decoded in many different ways, all equally "valid." However, under almost all of them, the resulting descriptions are not very complex or "interesting." For a description to be interesting, it must satisfy a *continuity requirement*, which is a refinement of a basic *reproducibility requirement.* We can express the latter by saying that if someone claims that a physical system S computes a function F, then for an arbitrary input x from the domain of F it must be possible to put S's input subsystem into an identifiable state decodable as x, and as a result S must evolve to a point where its output subsystem is in a state decodable as $F(x)$. The *piecewise continuity* requirement is that if it's possible only to get S's input subsystem into a state *close* to a previous state, that is, to a state decodable as $x' \approx x$, then except at a few boundaries the result is an output state decodable as $F(x') \approx F(x)$.[33]

Decodings are like frames of reference in physics, in that one must always be working within one, but the choice is arbitrary. The wrong choice of frame of reference can make a mathematical analysis much more difficult; the wrong choice of decoding can make computational analysis much less "interesting." But in each case the wrong choice does not produce *false* results; results from different frames do not contradict each other. A system

[33]Which is why John Searle has a large burden of proof in arguing his 1980 claim that a wall in his office is decodable as a word processor.

may compute one function under one decoding and a quite different one under another, but this is no endorsement of "observer-relativism" or any other sort of relativism, because one can use it to compute either function, whether or not anyone ever does use it, or notice that it can be used, that way. More to the point, if one finds a piece B of an animal's brain that computes different functions under different decodings, it is a further question which of these ever governs the behavior of the animal. It's possible that B is used to compute both functions, although only one at a time; or that B's results are sent to two different destinations, each employing a different decoding. However, I doubt these doublings ever happen,[34] because usually all but one decoding yield trivial computational descriptions.[35]

I have perhaps belabored the idea of decoding excessively, but I want to be clear that in exploring decodings we never get beyond the physical computational system; we are just redescribing it at an abstract level. The concept of *interpretation* is very different: An interpretation is a mapping from the domain of symbol structures to real-world entities they denote.[36] Let's assume for the sake of concreteness that I have a symbol in my head representing my first grandchild. I don't see the kid that often because he lives 5000 km away, but it seems correct to say that the symbol actually does stand for, or "denote," that particular toddler. What makes it the case that it does? Here the choice of mapping is not arbitrary, and all but one choice is wrong. So what singles out one interpretation as correct?

This is not the paper to try to provide an answer to this question. I will content myself with a few scattershot observations:

- We need a term for a "representation," absent the presupposition that it refers to or represents anything at all. Contra Haugeland, it is *not* a defining feature of such an entity that it come equipped with an

[34]One counterexample is the way DNA is read in many organisms, including humans. The same sequence of nucleotides can be read starting in different places, sometimes in different directions, yielding recipes for protein fragments that to be assembled in different ways (Lewis, 2001). Evolution is opportunistic, not elegant, so perhaps we should expect similar kludges in the brain.

[35]Of course, one can always indulge in tricks such as the following: scale the inputs and descale the outputs, so that under decoding D_1 B computes F_1, and under D_2 the input formerly decoded as x is decoded as $x' = 2x$ and the output formerly decoded as $F(x)$ is decoded as $F(x'/2)$. I don't doubt that we can avoid counting this infinitude of trivial notational variants of an"interesting" function by some technical maneuvers that are surely beside the point.

[36]Sometimes the term *model* is used as a synonym for "interpretation." I would prefer to reserve the term "model" (of a given set of belief representations) to mean "interpretation that makes that set *true*."

interpretation. I will adopt Newell and Simon's term *expression*, but do not assume that I mean Lisp expressions or the like. I recommend the same flexibility here as in the possible ranges of decodings. In fact, we can't escape a broad interpretation of "expression" because whether a physical state is to be thought of as a token of a given expression type is always relative to a decoding.

An expression consists of one or more symbols, arranged in linear sequences, or in more complex structures (trees, graphs, ...).

- At the very least, a claim that an expression represents a *changing* situation must be backed up by evidence that the expression changes to reflect the current state of the situation. (To be precise: under a fixed decoding the state changes to encode a different expression as the situation changes). Of course, there is always a time lag.

- It is not necessary to narrow down representations so they only occur in "minds," or organisms, or robots, or active agents of some kind. They can occur all over the place; they won't hurt anything.

- The temptation to differentiate "original" from "derived" meaning, as emphasized by Haugeland (AIVI, p. 119), must be resisted. Humans are supposed to be the wellspring of *intentionality,* the "aboutness" of thoughts and words, and our artifacts, such as books and DVDs, mean something only because we invented them specificially to mean those things. Or that's the idea anyway.

  The problem is that computers, which in normal use are just "super-books," whose intentionality is all derivative as far as this theory is concerned, seem to become a new source of the original sort of intentionality as they become more autonomous. It is hard to believe that a robot exploring an exoplanet on its own wouldn't be inventing its own new symbols to denote the species it discovered.

  As argued by Dennett (1987a), cognitive science is committed to analyzing and explaining human intentionality in the same terms it uses to explain the intentionality of autonomous robots. One can say that agent A can confer intentionality on a symbol system S by taking certain attitudes toward it, and possibly sharing it with other agents. S then has intentionality that is *derived from A,* or A's community, if the agent learned it from others. But whether one views A's intentionality as itself derived is another question.

- The notion of "intended" interpretation should play no role in psychosemantics. As explained in section 2.2, it is an error to assume continuity between the meanings of expressions in our heads and what we consciously *think* words and other expressions mean. Here as elsewhere, cognitive science should remain true to the strategy of finding a theory of unconscious thought first and then explaining conscious thought using that theory (Millikan, 1984, Dennett, 1991).

- Most thinkers about intentionality have supposed that the hardest part to explain is one's ability to think about nonexistent things such as Santa Claus, Sherlock Holmes, and the house one would like to build; and the way groups of people bring things into existence by believing in them, such as currency, corporations, and countries. But perhaps all of these can be assimilated to the case of fiction; and fiction can be explained as a story one is temporarily or permanently committed, to varying degrees.

- The elephant in the semantic room is the problem of indeterminacy. The classic source of anxiety on this head is Quine's (1960) theory of the "indeterminacy of translation" from one language to another, extended to all thought processes by Davidson (1973). The paradigmatic case in recent decades has been the "disjunction problem": if the symbol COW is tokened on some foggy evenings by horses, why don't we say it means "cow or horse seen on a foggy evening" (Fodor, 1988)?

I believe the "principle of charity" (Wilson, 1959, Field, 1978) can be adapted to solve this congeries of problems. This is the principle that we should interpret the words of an unfamiliar language so as to make as many of its speakers' utterances true as possible. Suppose we transpose this idea to representations in an agent's "brain" the goal being to make as many of its explicitly represented beliefs[37] true as possible. If we combine the principle of charity with ideas from minimum-description-length theory (Rissanen, 1989, Li and Vitányi, 2009), we might zero in on a single favored interpretation or a small neighborhood of interpretations.

I will say no more about this here, although there is much more to be said. I'll close this section with a couple of important reminders.

[37]Of course, representations are not tagged as being beliefs as opposed to desires, questions, or something else. To figure out which is which requires figuring out how the representations are connected to behavior and to transformations between categories (such as when *plans* are concocted to transform *desires* into *intentions*).

Those in pursuit of alternative "paradigms" for cognitive science often claim they can do without symbols (Shanahan, 2010) or that symbols only arise at the level of conscious thought, the vast unconscious being "sub-symbolic" (Smolensky, 1988). But symbols are like money. You can outlaw money, but something else will come to play the role of medium of exchange. You can outlaw symbols, but symbols might exist all the same. My favorite example is Damasio's theorizing about the unconscious mind in (Damasio, 1999). Much as he would like the brain to be explainable in terms of fairly clear ideas such as neural "maps," he must also acknowledge the existence of what he calls "images." The initial examples are reasonably clear:

> By *object* I mean entities as diverse as a person, a place, a melody, a toothache, a state of bliss; by *image* I mean a mental pattern in any of the sensory modalities, e.g., a sound image, a tactile image, the image of a state of well-being. Such images convey aspects of the physical characteristics of the object and they may also convey the reaction of like or dislike one may have for an object, the plans one may formulate for it, or the web of relationships of that object among other objects. (Damasio, 1999, p. 9)

The beginning of this paragraph combines our introspective understanding of the word "image" with the idea that the neural structures used in classifying visual, tactile, aural, or olfactory stimuli may be recruited when an image is activated (LeBihan et al., 1993, Kosslyn et al., 1995). But then images acquire property associations (likeability, e.g.), and become nodes in a "web of relationships" with other images. Most images are unconscious (Damasio, 1999, p.319). Eventually the properties and powers of images expand in indefinite ways:

> Images allow us to choose among repertoires of previously available patterns of action and optimize the delivery of the chosen action—we can, more or less deliberately, more or less automatically, review mentally the images which represent different options of action, different scenarios, different outcomes of action. We can pick and choose the most appropriate and reject the bad ones. Images also allow us to invent new actions to be applied to novel situations and to construct plans for future actions—the ability to transform and combine images of actions and scenarios is the wellspring of creativity. (Damasio, 1999, p. 23f)

By this point it is clear that most of the time most images are playing exactly the role symbols are supposed to play in computational theories of mind. My internal representation of my car must include detailed information about the visual patterns involved in recognizing it and the motor schemata used when unlocking it, sitting in it, and driving it. But when planning to use the car to run errands, the "image" of the car is basically used as a placeholder, a symbol, taking part in, e.g., representations of plans, but treated as having little internal structure itself.

So why not just acknowledge that the brain must contain symbols and expressions — representations in the computational sense?

> One ... reason to be cautious about the term representation is that it easily conjures up the metaphor of the brain as computer. The metaphor is inadequate, however. The brain does perform computations but its organization and working have little resemblance to the common notion of what a computer is. (Damasio, 1999, p. 321)

In other words, images must be "not-symbols" because people are "not-computers," anything but dreary old computers. This sort of reluctance to accept the ideas of symbol or computation, simply because they seem to entail that the brain is a programmable digital computer, is not unusual. Some responsibility for this tendency must lie with Haugeland, who provided a woolly and confused picture of AI that was hailed as a paragon of clarity. The typical reader of AIVI is hopelessly lost when it comes to understanding the differences between axiom systems, programmable digital computers, and computational models generally, so in rejecting the first two as a picture of the brain tends to reject the last as well, or to insist that programs are all right as long as they don't manipulate symbols, or to cling to Searle's 1980 distinction between "modeling" the mind and "being" a mind.

5 Whatever Happened to ...?

If Haugeland's exegesis of the early years of AI is not coherent, what shall we say about them? Was there an AI methodology that Haugeland just misdescribed? Or is there not much to say about whether there was a "mainstream" AI and if so what its core ideas were?

Something was different about those early years. I remember them well. The biggest difference between AI then and AI now is the increase in computing power by several orders of magnitude — but the impact is not what

one might think. If all you have is slow, tiny computers (measured by available memory), you tend to require every decision the computer makes to be "intelligent" in order for the outcome to look intelligent. So there was serious discussion about the problem with (e.g.) theorem provers being that they did search, but not "selective" search; their decisions were too "syntactic."

Having a zippy computer doesn't magically make life easy, as some popular accounts might lead one to believe. The first lesson of the computer age is that having a fast automated calculator does not cross off a category of problems; it just expands one's ambitions toward bigger problems of that category. What changed in my corner of the AI world is that if your program is making thousands or millions of decisions per second, then the idea goes by the boards that you need to think through each type of decision as if it were a conscious deliberation. Instead, like any other computer scientist, you start thinking about data structures, algorithms, and their time and space complexities. Of course, in the early 1970s these concepts were still in their infancy.

The "old-fashioned" way of thinking about how to program a computer is exemplified well by the way expert systems were built back in the 1980s.[38] In a process known as *knowledge acquisition* (Kendal and Creen, 2007), the programming team would meet with a group of experts in a domain such as medicine or computer configuration and try to write down rules that covered what the experts would do in every situation that could arise. As the program evolved, its recommendations would be compared with those of the experts, and discrepancies would be corrected by revising rules or adding new ones. But it would be a mistake to pick on this KA methodology, since it was basically the way everyone went about programming computers.[39]

The "computer as proxy" approach to programming can be contrasted with a modern approach in which masses of data are used to extract estimated probability distributions that are then fed into a Bayes-net solver to produce maximum-likelihood solutions to diagnostic or prediction problems (Manning and Schütze, 1999). I have no particular program in mind, but the paradigm, relying on "big data" as much as possible instead of experts' intuitions, is now becoming familiar to any reader of the *New York Times.*

There are several trends in AI that my hypothetical scenario exemplifies:

[38]Perhaps I should call them "GOFES."

[39]The obvious exception were the numerical analysts, who knew even before computation was fully automated that algorithms could not be relied on except in domains where they could be proven to be stable. Of course, we still use the "anthropomorphize and debug on corner cases" method in intro-programming courses.

- AI, which in its early years seemed wedded more to psychology than any other discipline, now teams with linguists, mathematicians, statisticans, mechanical engineers, neuroscientists, and anyone else whose ideas it can use. Meanwhile, psychologists have spurned AI for other suitors, especially neuroscience.

- Of all these disciplines, statistics has emerged as one of the cornerstones of AI, replacing logic, which aspired to be the theoretical framework of the discipline back in the 1970s. In those days uncertainty was the white whale that threatened to sink the logic ship, and how to represent and reason about it was a matter of considerable debate in the knowledge-representation (KR)[40] community. But following the work of Judea Pearl (1988) on *Bayes nets,* all doubts were cast aside and statistics, Bayesian and non-Bayesian, came to dominate as the favored technique for uncertain reasoning, and then for reasoning in general.

- In practical applications of knowledge representation, logic did not disappear but variables did. Propositional logic yields many tractable applications, and it interfaces well to statistics, since one can attach probabilities to symbols representing propositions. Quantified representations, including first-order logic, are more expressive, but no one really knows how to make them work with probability statements.[41] There have been advances in first-order theorem-proving algorithms, but they apply only to domains that can be characterized deductively using a fixed set of axioms.

- Programs are more rigorously tested than they used to be. In the 1960s and 70s a typical paper might show how a program performed on three or four examples (and display logs of its runs as if they were protocols from a subject in a psychology experiment). Nowadays a program must be tested on public repositories contain hundreds or thousands of problems in a standard notation, and its speed and coverage compared with those of previous programs on the same problem sets. Problems not already in the repository must be added to it so other experimenters can try their programs on them.

[40]This phrase is the standard way of referring to explicit representations of beliefs, theories, goals, plans, and other such things (Russell and Norvig, 2010). It has nothing to do with "knowledge" in the usual sense.

[41]But see Demey et al. (2014) for a review of proposals in this direction.

It is important to realize that even when logic seemed to be the theoretical armature of AI research, few practitioners used logic for much more than window dressing. One reason for this was that most inferences are not deductive. Another was that most programmers find it hard to think in terms of axioms. A programmer wants to make things *happen*, but logic is about timeless truth.[42] The Prolog language (Clocksin and Mellish, 2013), which is based on logic, requires many "extra-logical" enhancements to be usable for realistic applications. A good programmer must know when to switch from the purity of logic to the responsiveness of the enhancements.[43] I am less familiar with how programmers handle the tools available for statistical machine learning, but I would not be surprised if there was a similar disconnect between theory and practice here.

The fact is that there is no real computational theory of inference, intelligent or otherwise, that covers more than a few special cases. It may be that this is because it is still early days in the AI business, or for a deeper reason: that a bright line between inference and computation in general simply can't be drawn. Whenever a computation is a link in the causal chain from input to agent behavior, it is part of the "decision theory" for that agent. To the extent we can classify a class of representations as beliefs, the computations that lead to the creation of those representations can be classified as inferences. These computations may be describable as elegant algorithms, and there may be an explanation of why it's in the agent's interests to run those algorithms and store those representations, but there may simply be no underlying theory of inference from which these explanations are always drawn. Such a theory would explain why the agent should believe P at this time, and not believe Q, taking resource bounds into account. Logic has aspirations to be such a theoretical framework, as does decision theory, but for well known reasons they are fall short.

The one constant in the history of AI from 1970 to the present is that it was and remains empirical. Programs are judged by how well they work in practice, not by how well they work in theory. If you can prove that your program has an interesting property, that's great, but most of the time you can't. A chess program is good not because it can be proven to win, but because it *does* win. You may prove that it plays a certain endgame

[42]When someone says they "represented the knowledge procedurally," they mean, "I wrote a program."

[43]This is similar to knowing when to switch out of the purely functional subset of a language like Lisp (Seibel, 2005) or Scala (Horstmann, 2012), with the added complexity of Prolog's nondeterminism. It takes years to learn to navigate these waters, and most programmers never do.

perfectly, but that's not usually the central point of the work. The difference between 1970 and now is that the empirical results are more convincing.[44]

6 Conclusion: GOFAI, the Very Idea

Haugeland's notion of "good old-fashioned artificial intelligence" does not withstand close scrutiny. It overemphasizes the idea of "interpreted formal system," leaving us baffled as to how it applies to real computer programs, especially those of Type B, which do not contain explicit formal systems at all. The contemporaneous "physical-system hypothesis" of Newell and Simon (1976) is even less coherent or plausible.

The influence of Haugeland's analysis and that of Newell and Simon has been great, measured in citations, but no one really knows what they are citing. They all describe it in quite different terms. It would be simplistic to blame all confusion about AI methodology on one or two papers or people. But the papers I've talked about seem to be the nucleus of the literature that is referred to whenever someone summarizes (early) AI research. It is not hard to see why AIVI would play such a central role. It is so well written, and so lucid in places, that at first glance it seems to put a particular subculture or phase of artificial intelligence under a microscope. But when you try to make sense of the resulting image, there are sharp fractures visible in the lenses.

I've coined the word *symbolicism* for the (alleged) theory described by Haugeland, and Newell and Simon. Inevitably, there is little agreement about what symbolicism comes to. It's mainly useful for setting up AI as a target. If someone thinks human thought is basically X, and that symbolicism isn't X, then AI, at least the symbolicist sort, has no place in an enlightened cognitive science. X is typically "general-purpose," "insightful," "creative," "adaptable," "inconsistency-tolerant," "learning-based," "fast," "massively parallel," "opaque," or "transparent."

The collaboration between AI and cognitive psychology has diminished sharply over the last few decades. As each field sought to make use of more complex models, they found it harder to communicate. Psychologists can't run experiments if they have to build working computational models along the way, because building such models takes too long. Even apparent counterexamples such as ACT-R (Anderson et al., 2004) depend on simulated vision modules for crucial pieces of information from the environment. Meanwhile, many of the ideas making their way into AI seemed to have little

[44]Russell and Norvig (2010, pp.16–28) give a concise and useful history of AI.

to do with psychology. When the Watson program won a simplified version of Jeopardy in 2011, it relied on searching a large database of stored text articles, then using Bayesian methods to reject most of the possible matches. It would be surprising if good human jeopardy players' brains worked this way.

One reason psychologists have given up on detailed computational models is that AI has been slow to deliver them. ACT-R models depend on simulated vision for an excellent reason: there is no general-purpose computer-vision module to be had that performs anywhere near as well as the human visual system. The situation with respect to language is just as bad. Although speech-recognition systems are approaching human performance (Saon and Chien, 2012), linguists seem to be getting further and further from consensus on the grammar of even a single human language. Nor has computational linguistics, at the interface beween AI and linguistics, delivered tidy solutions that psychologists can use.

Philosophers have too often played Iago in the story of divorce between AI and psychology. Haugeland's story and the rise of connectionism provided a theory of history that replaces the need to follow the actual history of AI. Philosophers read and dissected the early AI papers, but how many have tried to read papers on "deep learning" (Hinton, 2007), one of the current trends in neural-net theory?

I can only conjecture that the lack of interest is partly because the math has gotten tougher in the last 50 years, and partly because philosophers don't *want* the brain to be analyzable, once materialism and perhaps functionalism have been genuflected to. I apologize for this *ad hominem* argument; I raise it only to point out that philosophers' fears, if they are prone to them, are perfectly understandable. Nobody wants to be analyzable, including me. But there's no reason to worry that scientific explanations of brains will explain *us* as individuals. Suppose there is eventually a computational explanation of, say, love. It would suggest why different people have different "types" they are likely to fall for, the sorts of "learning" that causes people to change their attitudes toward lovers over time, and so forth. If you are willing to submit to analysis of your brain, predictions can be made of how and why you fall in love, and possibly stay in love, with whom you do. Well, what difference would this make to *you*? The way you think about love may be inaccurate, but simply knowing this fact would not be enough to change the way you navigate through your love life (McDermott, 2013). If we use the phrase "theory of love" for the way you think about love (by analogy with "theory of mind"), just knowing how your brain controls your love life

wouldn't allow you to "escape" the theory (or make you want to escape).[45]

Nonetheless, philosophers skeptical about the prospects for a computational theory of the brain and mind can point to the failure to achieve a blitzkrieg victory over the problem as justifying rejection of the whole cognitive-scientific enterprise. Exhibit A: Colin McGinn (2013, online) —

> Even in sober neuroscience textbooks we are routinely told that bits of the brain "process information," "send signals," and "receive messages"—as if this were as uncontroversial as electrical and chemical processes occurring in the brain. We need to scrutinize such talk with care. Why exactly is it thought that the brain can be described in these ways? It is a collection of biological cells like any bodily organ, much like the liver or the heart, which are not apt to be described in informational terms. It can hardly be claimed that we have observed information transmission in the brain, as we have observed certain chemicals; this is a purely theoretical description of what is going on. ...
>
> ... Why do we say that telephone lines convey information? Not because they are intrinsically informational, but because conscious subjects are at either end of them, exchanging information in the ordinary sense. Without the conscious subjects and their informational states, wires and neurons would not warrant being described in informational terms.
>
> ... It is simply false to say that one neuron literally "sends a signal" to another; what it does is engage in certain chemical and electrical activities that are causally connected to genuine informational activities [occurring in the mind].

This breathtaking repudiation of the metaphors underlying communication theory, when they are applied to biological systems, leaves little hope for the rest of cognitive science. Granted, McGinn and his fellow "reactionaries," such as John Searle (1992), represent an extreme within philosophy of mind. But the high regard in which they are held has encouraged a retreat from computational theorizing about the mind, often all the way back to some form of dualism.[46]

[45]Sellars made a distinction between the "scientific" and "manifest" images of reality, and of human doings in particular in (Sellars, 1962). In those terms, few people succeed for a large fraction of their time in behaving as though the scientific image were the reality. See also McDermott (2013).

[46]I should note that both McGinn and Searle claim to be materialists, but neither

The collaboration between AI and other branches of cognitive science has never been extinguished completely, of course. Bright spots include the conference and journal Biologically Inspired Cognitive Architectures (BICA), the International Conference on Cognitive Modeling (ICCM), and of course the journal *Cognitive Science*, which was present at the creation. Perhaps they will provide the seeds of a renaissance.

I apologize if my remarks on alternatives to Haugeland's ideas are simultaneously too brief to be intelligible, and too numerous to skim quickly. I hope to amplify them in future publications, especially in the area of "robosemantics."

Acknowledgements: Thanks to Frank Keil and Laurie Santos for explaining why "theory of mind" is not regarded by most psychologists as an explicitly represented theory, and more generally what role representation plays in the current methodology of cognitive psychology. Thanks to Ken Forbus for feedback on an earlier draft, and for his views on the future of symbolic AI and its collaboration with psychology. None of these people is to blame for my interpretation of their suggestions.

References

Agre, P. E. (1997). *Computation and Human Experience*. Cambridge University Press.

Alter, T. and Walter, S. (2006). *Phenomenal Concepts and Phenomenal Knowledge:New Essays on Consciousness and Physicalism*. Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S. A., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psych. Rev*, 111(4):1036–1060.

Andrew, A. (1987). Review of Haugeland, *artificial intelligence: the very idea*. *Robotica*, 5(2):137.

Aue, T., Lavelle, L. A., and Cacioppo, J. T. (2009). Great expectations: What can fMRI research tell us about psychological phenomena? *Int. J. of Psychophysiology*, 73(1):10–16.

acknowledges the slightest bit of progress in explaining the conscious mind, the only mind they recognize, in physicalistic terms; McGinn believes there never will be (McGinn, 1991).

Backus, J. W., Beebert, R. J., Best, S., Goldberg, R., Haibt, L. M., Herrick, H. L., Nelson, R. A., Sayre, D., Sheridan, P. B., Stern, H., Ziller, I., Hughes, R. A., and Nutt, R. (1957). The FORTRAN automatic coding system. In *Proc. Western Joint Computer Conference*, pages 188–198.

Baumgartner, P. and Payr, S., editors (1995). *Speaking Minds: Interviews with Twenty Eminent Cognitive Scientists*. Princeton University Press.

Bickhard, M. H. and Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*. North-Holland, Amsterdam.

Botterill, G. and Carruthers, P. (1999). *The Philosophy of Psychology*. Cambridge University Press.

Chalmers, D. (1994). On implementing a computation. *Minds and Machines*, 4(4):391–402.

Chemero, A. (2013). Radical embodied cognitive science. *Review of Gen. Psych*, 17(2):145–150.

Clocksin, W. and Mellish, C. (2013). *Programming in Prolog: Using the ISO Standard*. Springer, New York.

Colodny, R. G., editor (1962). *Frontiers of Science and Philosophy*. University of Pittsburgh Press.

Copeland, B. J., editor (2005). *Alan Turing's Automatic Computing Engine: The Master Codebreaker's Struggle to Build the Modern Computer*. Oxford University Press.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Bradford Books/MIT Press.

Cummins, R. and Schwarz, G. (1987). Radical connectionism. *Southern J. of Phil*, 26(Supplement):43–61.

Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt, Inc, San Diego. (A Harvest Book.).

Davidson, D. (1973). Radical interpretation. *Dialectica*, 27(3-4):313–328.

Demey, L., Kooi, B., and Sack, J. (2014). Logic and probability. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Online at . , at.

Dennett, D. C. (1987a). Evolution, error, and intentionality. In Dennett (1987b), pages 287–321.

Dennett, D. C. (1987b). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company, Boston.

Dennett, D. C. (2006). What RoboMary knows. In Alter and Walter (2006), pages 15–31.

Dijkstra, E. (1968). Go To statement considered harmful. *Comm. ACM*, 11(3):147–48.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press, Cambridge, Mass.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. MIT Press, Cambridge, Mass.

Dreyfus, H. (1972). *What Computers Can't Do*. Harper & Row, New York.

Dreyfus, H. L. (2007). Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Artificial Intelligence*, 171:1137–1160.

Dreyfus, H. L. and Dreyfus, S. E. (1988). Making a mind versus modelling the brain: artificial intelligence back at the branchpoint. *Daedalus*, 117(1):15–43. Reprinted in (Negrotti 1991), pp. 33–54; page references are to this edition.

Eckardt, B. V. (1988). Review of Haugeland, *artificial intelligence: the very idea*. *The Philosophical Review*, 97(2):286–290.

Field, H. (1978). Mental representation. *Erkenntnis*, 13(1):9–61.

Fikes, R., Hart, P. E., and Nilsson, N. J. (1972). Learning and executing generalized robot plans. *Artificial Intelligence*, 3(4):349–371.

Flanagan, O. (1986). Review of Haugeland, *artificial intelligence: the very idea*. *Philosophy in Review*, 6(10):474–76.

Fodor, J. (1975). *The Language of Thought*. Thomas Y. Crowell, New York.

Fodor, J. (1978). Tom swift and his procedural grandmother. *Cognition*, 6:204–224. Also in Fodor 1981.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press, Cambridge, Mass.

Fodor, J. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In Pylyshyn (1987).

Fodor, J. (1988). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Bradford Books/MIT Press, Cambridge, Mass.

Fodor, J. (1990). *A Theory of Content and Other Essays*. Bradford Books/MIT Press, Cambridge, Mass.

Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press, Oxford.

Forbus, K. D. (2010). *AI and cognitive science: the past and next 30 years. Topics in Cognitive Science* **2**.

Ford, N. (1987). Review of Haugeland, *artificial intelligence: the very idea. International J. of Information Management*, 7(1):59–60.

Gallistel, C. R. and King, A. P. (2009). *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. Wiley/Blackwell, New York.

Gentner, D. and Forbus, K. D. (2011). Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):266–276.

Gentner, D., Holyoak, K. J., and Kokinov, B. K., editors (2001). *The Analogical Mind: Perspectives from Cognitive Science*. The MIT Press, Cambrdige, Mass.

Goldman, A. I. (2012). Theory of mind. *Oxford Handbook of Philosophy of Cognitive Science*, pages 402–424.

Gopnik, A. and Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind and Language*, 7:145–171.

Graziano, M. S. A. and Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, 2(2):98–113.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2):389–400.

Haidt, J. and Joseph, C. (2004). Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Harnad, S. (2001). Grounding symbols in the analog world with neural nets – a hybrid model. *Psycoloquy*, 12(34).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York. Second edition.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences*, 2:215–226.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge, Mass.

Haugeland, J. (1990). The intentionality all-stars. *Phil. Perspectives*, 4:383–427. (Action Theory and Philosophy of Mind).

Hinsley, F. and Stripp, A., editors (1993). *Codebreakers: The inside story of Bletchley Park*. Oxford University Press, Oxford. (ISBN 978-0-19-280132-6).

Hinton, G. E. (2007). Learning multiple layers of representation. *TRENDS in Cognitive Sciences*, 11(10):428–434.

Horstmann, C. S. (2012). *Scala for the Impatient*. Addison-Wesley, New Jersey. Upper Saddle River.

Itkonen, E. (2015). *Analogy as structure and process: Approaches in linguistics, cognitive psychology and philosophy of science*. John Benjamins Publishing.

Jacquette, D. (2009). *The Philosophy of Mind: The Metaphysics of Consciousness*. Continuum, London. (Note: Revised and expanded version of Jacquette 1994. Full name of publisher is Continuum International Publishing Group.).

Kendal, S. and Creen, M. (2007). *An Introduction to Knowledge Engineering*. Springer, London.

Kosslyn, S., Thompson, W., Kim, I., and Alpert, N. (1995). Topographical representations of mental images in primary visual cortex. *Nature*, 378(6556):496–98.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

LeBihan, D., Turner, R., Zeffiro, T. A., Cuénod, C. A., Jezzard, P., and Bonnerot, V. (1993). Activation of human primary visual cortex during visual recall: A magnetic resonance imaging study. In *Proc. Natl. Acad. Sci*, volume 90, pages 11802–805.

Lewis, R. (2001). Genome economy. *The Scientist*, 15(12):19–21.

Li, M. and Vitányi, P. M. (2009). *An Introduction to Kolmogorov Complexity And Its Applications*. Springer Science & Business Media.

Lucas, J. (1961). Minds, machines and goedel. *Philosophy*, 36:112–127. Reprinted in (Anderson 1964), pp. 43–59.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Marr, D. (1982). *Vision*. Freeman. W.H.

McDermott, D. (1987). We've been framed, or, why AI is innocent of the frame problem. In Pylyshyn (1987).

McDermott, D. (2001a). Digital computers as red herrings. *Psycoloquy*, 12(54). Commentary on (Harnad 2001).

McDermott, D. (2001b). *Mind and Mechanism*. MIT Press, Cambridge, Mass.

McDermott, D. (2013). Computationally constrained beliefs. *J. of Consciousness Studies*, 20(5–6):124–50.

McGinn, C. (1991). *The Problem of Consciousness*. Basil Blackwell, Oxford.

McGinn, C. (2013). Homunculism. *New York Review of Books,*. March 21, 2013.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. The MIT Press, Cambridge, MA.

Millikan, R. G. (1984). *Language, Thought, and other Biological Categories.* MIT Press, Cambridge, MA.

Morris, C. (1993). Navy Ultra's poor relations. In Hinsley and Stripp (1993), pages 231–245. (ISBN 978-0-19-280132-6).

Neander, K. (1988). Review of Haugeland, *artificial intelligence: the very idea. Australasian J. of Philosophy*, 66(2):269–71.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4:135–183.

Newell, A. (1994). *Unified Theories of Cognition.* Harvard University Press.

Newell, A. and Shaw, J. (1957). Programming the Logic Theory Machine. In *Proc. Western Computer Conf*, pages 230–240.

Newell, A., Shaw, J., and Simon, H. A. (1957). Empirical explorations of the Logic Theory Machine: a case study in heuristic. In *Proc. Western Computer Conf*, pages 218–230.

Newell, A., Shaw, J. C., and Simon, H. A. (1959). Report on a general problem-solving program. In *Proc. IFIP Congress*, pages 256–264.

Newell, A. and Simon, H. (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A. and Simon, H. (1976). Computer science as empirical inquiry: symbols and search. *Comm. ACM*, 19(3):113–126.

Nowachek, M. T. (2014). Why robots cant become racist, and why humans can. *PhaenEx*, 9(1):57–88.

Ogilvie, J. W. (1986). Review of Haugeland, *artificial intelligence: the very idea. AI Magazine*, 7(1):86–87.

Patterson, D. A. and Hennessy, J. L. (2009). *Computer Organization and Design: The Hardware/Software Interface (4th Edition).* Morgan Kaufmann, Burlington, Massachusetts.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, Los Altos, CA.

Pohl, I. (1988). Review of Haugeland, *artificial intelligence: the very idea. J. of Symbolic Logic*, 53(2):659–660.

Pope, M. (1999). *The Story of Decipherment.* Thames and Hudson, London.

Pylyshyn, Z. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3:111–169.

Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science.* MIT Press, Cambridge.

Pylyshyn, Z., editor (1987). *The Robot's Dilemma: The Frame Problem and Other Problems of Holism in Artificial Intelligence.* Ablex Publishing Co, Norwood, N.J.

Quine, W. V. (1960). *Word and Object.* Wiley, New York.

Raichle, M. E. (1999). Modern phrenology: maps of human cortical function. *Annals of the New York Academy of Sciences*, 882(1).

Rey, G. (1997). *Contemporary Philosophy of Mind: A Contentiously Classical Approach.* Blackwell Publishers, Cambridge, Mass.

Richmond, S. (1991). Review of Haugeland, *artificial intelligence: the very idea. Computers and the Humanities*, 25(5):331–37.

Rissanen, J. (1989). *Statistical Complexity in Statistical Inquiry.* World Scientific, Singapore.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, Mass.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (3rd edition).* Prentice Hall.

Saon, G. and Chien, J.-T. (2012). Large-vocabulary continuous speech recognition systems: A look at some recent advances. *IEEE Signal Processing Magazine*, 29(6):18–33.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3:417–424.

Searle, J. R. (1990). Is the brain's mind a computer program? *Scientific American*, 262:26–31.

Searle, J. R. (1992). *The Rediscovery of the Mind.* MIT Press, Cambridge, Mass.

Seibel, P. (2005). *Practical Common Lisp*. Apress.

Sellars, W. (1962). Philosophy and the scientific image of man. In Colodny (1962), pages 35–78. (Reprinted in (Sellars 1963), pp. 1–40.).

Shanahan, M. (2010). *Embodiment and the Inner Life*. Oxford University Press.

Smith, B. C. (1988). *The Semantics of Clocks*. Springer Netherlands.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74.

Spivey, M. (2007). *The Continuity of Mind*. Oxford University Press.

Turing, A. (1936). On computable numbers, with an application to the *entscheidungsproblem*. In *Proc. London Math. Society (2)* , volume 42, pages 230–265. (Correction 1937 vol **43**, pp. 544-546).

van Gelder, T. (1998). The dynamic hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21:615–665. (With commentary and rebuttal).

Vellino, A. (1986). Review of Haugeland, *artificial intelligence: the very idea*. *Artificial Intelligence*, 29:349–53.

Wikipedia (2015). PDP-10.

Wilson, N. L. (1959). Substances without substrata. *The Review of Metaphysics*, 12(4):521–539.

Winograd, T. and Flores, F. (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Ablex Publishing Corporation, New Jersey. Norwood.