Artificial Intelligence and Consciousness

Drew McDermott
Yale University

This paper is essentially the same as that published as chapter 6 (pages 117–150) of Philip David Zelazo, Morris Moscovitch, and Evan Thompson (eds.) 2007 *The Cambridge Handbook of Consciousness.* Cambridge University Press

Abstract: Consciousness is only marginally relevant to artificial intelligence (AI), because to most researchers in the field other problems seem more pressing. However, there have been proposals for how consciousness would be accounted for in a complete computational theory of the mind, from theorists such as Dennett, Hofstadter, McCarthy, McDermott, Minsky, Perlis, Sloman, and Smith. One can extract from these speculations a sketch of a theoretical synthesis, according to which consciousness is the property a system has by virtue of modeling itself as having sensations and making free decisions. Critics such as Harnad and Searle have not succeeded in demolishing a priori this or any other computational theory, but no such theory can be verified or refuted until and unless AI is successful in finding computational solutions of difficult problems such as vision, language, and locomotion.

1 Introduction

Computationalism is the theory that the human brain is essentially a computer, although presumably not a stored-program, digital computer, like the kind Intel makes. *Artificial intelligence* (AI) is a field of computer science that explores computational models of problem solving, where the problems to be solved are of the complexity of problems solved by human beings. An AI researcher need not be a computationalist, because they[1] might believe that computers can do things brains do noncomputationally. Most AI researchers are computationalists to some extent, even if they think digital computers and brains-as-computers compute things in different ways. When it comes to the problem of phenomenal consciousness, however, the AI researchers who care about the problem and believe that AI can solve it are a tiny minority, as we will see. Nonetheless, because I count myself in that minority, I will do my best to survey the work of my fellows and defend a version of the theory that I think represents that work fairly well.

Perhaps calling computationalism a "theory" is not exactly right here. One might prefer "working hypothesis," "assumption," or "dogma." The evidence for computationalism is not overwhelming, and some even believe it has been refuted, by a priori arguments or empirical evidence. But, in some form or other, the computationalist hypothesis underlies modern research in cognitive psychology, linguistics, and some kinds of neuroscience. That is, there

[1] To avoid sexist pronouns, I will sometimes use third-person-plural pronouns to refer to a generic person.

wouldn't be much point in considering formal or computational models of mind if it turned out that most of what the brain does is not computation at all, but, say, some quantum-mechanical manipulation (Penrose, 1989). Computationalism has proven to be a fertile working hypothesis, although those who reject it typically think of the fertility as similar to that of fungi, or of pod people from outer space.

Some computationalist researchers believe that the brain is nothing more than a computer. Many others are more cautious, and distinguish between modules that are quite likely to be purely computational (e.g., the vision system), and others that are less likely, such as the modules, or principles of brain organization, that are responsible for creativity, or romantic love. There's no need, in their view, to require that absolutely everything be explained in terms of computation. The brain could do some things computationally and other things by different means, but if the parts or aspects of the brain that are responsible for these various tasks are more or less decoupled, we could gain significant insight into the pieces that computational models are good for, and leave the other pieces to some other disciplines such as philosophy and theology.[2]

Perhaps the aspect of the brain that is most likely to be exempt from the computationalist hypothesis is its ability to produce consciousness, that is, to experience things. There are many different meanings of the word "conscious," but I am talking here about the "Hard Problem" (Chalmers, 1996), the problem of explaining how it is that a physical system can have vivid experiences with seemingly intrinsic "qualities," such as the redness of a tomato, or the spiciness of a taco. These qualities usually go by their Latin name, *qualia*. We all know what we're talking about when we talk about sensations, but they are notoriously undefinable. We all learn to attach a label such as "spicy" to certain tastes, but we really have no idea whether the sensation of spiciness to me is the same as the sensation of spiciness to you.

Perhaps tacos produce my "sourness" in you, and lemons produce my "spiciness" in you.[3] We would never know, because you have learned to associate the label "sour" with the quale of the experience you have when you eat lemons, which just happens to be very similar to the quale of the experience I have when I eat tacos. We can't just tell each other what these qualia are like; the best we can do is talk about comparisons. But we agree on questions such as, Do tacos taste more like Szechuan chicken or more like lemons?

I focus on this problem because other aspects of consciousness raise no special problem for computationalism, as opposed to cognitive science generally. The purpose of consciousness, from an evolutionary perspective, is often held to have something to do with allocation and organization of scarce cognitive resources. For a mental entity to be conscious is for it to be

[2] I would be tempted to say there is a spectrum from "weak" to "strong" computationalism to reflect the different stances on these issues, but the terms "weak" and "strong" have been used by John Searle (1980) in a quite different way. See section 5.2.

[3] I am taking this possibility seriously for now because everyone will recognize the issue and its relationship to the nature of qualia. But I follow Sloman & Chrisley (2003) in believing that cross-personal comparison of qualia makes no sense. See section 3.4 and (McDermott, 2001), ch. 4.

held in some globally accessible area (Baars, 1988, 1997). AI has made contributions to this idea, in the form of specific ideas about how this global access works, going under names such as the "blackboard model" (Hayes-Roth, 1985), or "agenda-based control" (Currie & Tate, 1991). One can evaluate these proposals by measuring how well they work, or how well they match human behavior. But there doesn't seem to be any *philosophical* problem associated with them.

For phenomenal consciousness, the situation is very different. Computationalism seems to have nothing to say about it, simply because computers don't have experiences. I can build an elaborate digital climate-control system for my house, which keeps its occupants at a comfortable temperature, but the climate-control system never feels overheated or chilly. Various physical mechanisms implement its temperature sensors in various rooms. These sensors produce signals that go to units that compute whether to turn the furnace on or turn the air conditioner on. The result of these computations cause switches to close so that the furnace or air conditioner does actually change state. We can see the whole path from temperature sensing to turning off the furnace. Every step can be seen to be one of a series of straightforward physical events. Nowhere are you tempted to invoke conscious sensation as an effect or element of the causal chain.

This is the prima facie case against computationalism, and a solid one it seems to be. The rest of this article is an attempt to dismantle it.

2 An Informal Survey

Although one might expect AI researchers to adopt a computationalist position on most issues, they tend to shy away from questions about consciousness. AI has often been accused of being over-hyped, and the only way to avoid the accusation, apparently, is to be so boring that journalists stay away from you. As the field has matured, and as a flock of technical problems have become its focus, it has become easier to bore journalists. The last thing most serious researchers want is to be quoted on the subject of computation and consciousness.

In order to get some kind of indication of what positions researchers take on this issue, I conducted an informal survey of Fellows of the American Association for Artificial Intelligence in the summer of 2003. I sent e-mail to all of them asking the following question:

> Most of the time AI researchers don't concern themselves with philosophical questions, as a matter of methodology and perhaps also opinion about what is ultimately at stake. However, I would like to find out how the leaders of our field view the following problem: Create a computer or program that has "phenomenal consciousness," that is, the ability to experience things. By "experience" here I mean "qualitative experience," the kind in which the things one senses seem to have a definite but indescribable quality, the canonical example being "looking red" as opposed to "looking green."
>
> Anyway, please choose from the following possible resolutions of this problem:
>
> 1. The problem is just too uninteresting compared to other challenges

1 Problem uninteresting	3%	
2a Ill-defined	11%	19%
2b Only apparent	8%	
3 AI silent	7%	
4 Requires new ideas	32%	
5 AI will solve it as is	3%	
6 Solution in sight	15%	
7 None of the above	21%	

Percentages indicate fraction of the 34 who responded

Table 1: Results of survey of AAAI Fellows

2. The problem is too ill defined to be interesting; or, the problem is only apparent, and requires no solution
3. It's an interesting problem, but AI has nothing to say about it
4. AI researchers may eventually solve it, but will require new ideas
5. AI researchers will probably solve it, using existing ideas
6. AI's current ideas provide at least the outline of a solution
7. My answer is not in the list above. Here it is:...

> Of course, I don't mean to exclude other branches of cognitive science; when I say "AI" I mean "AI, in conjunction with other relevant disciplines." However, if you think neuroscientists will figure out phenomenal consciousness, and that their solution will entail that anything not made out of neurons cannot possibly be conscious, then choose option 3.
>
> Because this topic is of passionate interest to a minority, and quickly becomes annoying to many others, please direct all followup discussion to fellows-discuss@aaai.org. Directions for subscribing to this mailing list are as follows: ...
>
> Thanks for your time and attention.

Of the approximately 207 living Fellows, I got responses from 34. The results are as indicated in Table 1.

Of those who chose 7 (None of the above) as answer, here are some of the reasons why:

> "Developing an understanding of the basis for conscious experience is a central, long-term challenge for AI and related disciplines. It's unclear at the present time whether new ideas will be needed...."

> “If two brains have isomorphic computation then the ‘qualia’ must be the same. Qualia must be just another aspect of computation — whatever we say of qualia must be a property of the computation viewed as computation.”
>
> “There are two possible ways (at least) of solving the problem of phenomenal consciousness, ‘explaining what consciousness is’ and ‘explaining consciousness away.’ It sounds like you are looking for a solution of the first type, but I believe the ultimate solution will be of the second type.”
>
> “The problem is ill-defined, and always will be, but this does not make it uninteresting. AI will play a major role in solving it.”

If the table seems to indicate no particular pattern, just remember that what the data show is that the overwhelming majority (173 out of 207) refused to answer the question at all. Obviously, this was not a scientific survey, and the fact that its target group contained a disproportionate number of Americans perhaps biased it in some way. Furthermore, the detailed responses to my questions indicated that respondents understood the terms used in many different ways. But if 84% of AI Fellows don’t want to answer, we can infer that the questions are pretty far from those that normally interest them. Even the 34 who answered include very few optimists (if we lump categories 5 and 6 together), although about the same number (categories 1 and 2) thought the problem didn’t really need to be solved. Still, the outright pessimists (category 3) were definitely in the minority.

3 Research on Computational Models of Consciousness

In view of the shyness about consciousness shown by serious AI researchers, it is not surprising that detailed proposals about phenomenal consciousness from this group should be few and far between.

3.1 Moore/Turing Inevitability

One class of proposals can be dealt with fairly quickly. Hans Moravec, in a series of books (Moravec, 1988, 1999), and Raymond Kurzweil (1999) have more or less assumed that continuing progress in the development of faster, more capable computers will cause computers to equal and then surpass humans in intelligence, and that consciousness will be an inevitable consequence. The only argument offered is that the computers will talk as though they are conscious; what more could we ask?

I believe a careful statement of the argument might go like this:

1. Computers are getting more and more powerful.

2. This growing power allows computers to do tasks that would have been considered infeasible just a few years ago. It is reasonable to suppose, therefore, that many things we think of as infeasible will eventually be done by computers.

3. Pick a set of abilities such that if a system had them we would deal with it as we would a person. The ability to carry on a conversation must be in the set, but we can imagine lots of other abilities as well: skill in chess, agility in motion, visual perspicacity, and so forth. If we had a talking robot that could play poker well, we would treat it the same way we treated any real human seated at the same table.

4. We would feel an overwhelming impulse to attribute consciousness to such a robot. If it acted sad at losing money, or made whimpering sounds when it was damaged, we would respond as we would to a human that was sad or in pain.

5. This kind of overwhelming impulse is our only evidence that a creature is conscious. In particular, it's the only real way we can tell that *people* are conscious. Therefore, our evidence that the robot was conscious would be as good as one could have. Therefore the robot would *be* conscious, or be conscious for all intents and purposes.

I call this the "Moore/Turing inevitability" argument because it relies on Moore's Law (Moore, 1965) predicting exponential progress in the power of computers, plus a prediction about how well future programs will do on the "Turing test," proposed by Alan Turing (1950) as a tool for rating how intelligent a computer is.[4] Turing thought all questions about the *actual* intelligence (and presumably degree of consciousness) of a computer were too vague or mysterious to answer. He suggested a behaviorist alternative: Let the computer carry on a conversation over a teletype line (or via an instant-messaging system, we would say today). If a savvy human judge could not distinguish the computer's conversational abilities from those of a real person at a rate better than chance, then we would have some measure of the computer's intelligence. We could use this measure *instead of* insisting on measuring the computer's *real* intelligence, or *actual* consciousness.

This argument has a certain appeal. It certainly seems that *if* technology brings us robots that we can't help treating as conscious, then in the argument about whether they really are conscious the burden of proof will shift, in the public mind, to the party-poopers who deny that they are. But so what? You can't win an argument by imagining a world in which you've won it and declaring it inevitable.

The anti-computationalists can make several plausible objections to the behavioral-inevitability argument:

[4] Turing actually proposed a somewhat different test. See (Davidson, 1990) for discussion. Nowadays this version is the one everyone works with.

- Just because computers have made impressive strides doesn't mean that *any* task we set them they will eventually be able to carry out. In particular, progress in carrying on conversations has been dismal.[5]

- Even if a computer could carry on a conversation, that wouldn't tell us *anything* about whether it really was conscious.

- Overwhelming impulses are not good indicators for whether something is true. The majority of people have an overwhelming impulse to believe that there is such a thing as luck, so that a lucky person has a greater chance of winning at roulette than an unlucky person. The whole gambling industry is based on exploiting the fact that this absurd theory is so widely believed.

I will come back to the second of these objections in section 5.1. The others I am inclined to agree with.

3.2 Hofstadter, Minsky, McCarthy

Richard Hofstadter touches on the problem of consciousness in many of his writings, especially the material he contributed to (Hofstadter & Dennett, 1981). Most of he what he writes seems to be intended to stimulate or tantalize one's thinking about the problem. For example, in (Hofstadter, 1979) there is a chapter (reprinted in (Hofstadter & Dennett, 1981)) in which characters talk to an anthill. The anthill is able to carry on a conversation because the ants that compose it play roughly the role neurons play in a brain. Putting the discussion in the form of a vignette allows for playful digressions on various subjects. For example, the anthill offers the anteater (one of the discussants) some of its ants, which makes vivid the possibility that "neurons" could implement a negotiation that ends in their own demise.

It seems clear reading the story that Hofstadter believes that the anthill is conscious, and therefore one could use integrated circuits rather than ants to achieve the same end. But most of the details are left out. In this as in other works, it's as if he wants to invent a new, playful style of argumentation, in which concepts are broken up and tossed together into so many configurations that the original questions one might have asked get shunted aside. If you're already convinced by the computational story, then this conceptual play is delightful. If you're a skeptic, I expect it can get a bit irritating.

I put Marvin Minsky in this category as well; perhaps it should be called "Those who don't take consciousness very seriously as a problem." He wrote a paper in 1968 (Minsky, 1968b) that introduced the concept of *self-model*, which, as we will see, is central to the computational theory of consciousness.

[5] The Loebner Prize is awarded every year to the writer of a program that appears "most human" to a panel of judges. You can see how close the programs are getting to fooling anyone at the website, `http://www.loebner.net/Prizef/loebner-prize.html` .

> To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A.. . . If A is the world, questions for A are experiments. A* is a good model of A, in B's view, to the extent that A*'s answers agree with those of A, on the whole, with respect to the questions important to B. When a man M answers questions about the world, then (taking on ourselves the role of B) we attribute this ability to some internal mechanism W* inside M.

This part is presumably uncontroversial. But what's interesting is that W*, however it appears, will include a model of M himself, M*. In principle, M* will contain a model of W*, which we can call W**. M can use W** to answer questions about the way he (M) models the world. One would think that M** (the model of M* in W**) would be used to answer questions about the way M models himself, but Minsky has a somewhat different take: M** is used to

> answer general questions about himself. Ordinary questions about himself, e.g., how tall he is, are answered by M*, but very broad questions about his nature, e.g., what kind of a thing he is, etc., are answered, if at all, by descriptive statements made by M** about M*.

Now, the key point is that the accuracy of M* and M** need not be perfect.

> A man's model of the world has a distinctly bipartite structure: One part is concerned with matters of mechanical, geometrical, physical character, while the other is associated with things like goals, meanings, social matters, and the like. This division of W* carries through the representations of many things in W*, especially to M itself. Hence, a man's model of himself is bipartite, one part concerning his body as a physical object and the other accounting for his social and psychological experience.

This is why dualism is so compelling. In particular, Minsky accounts for free will by supposing that it develops from a "strong primitive defense mechanism" to resist or deny compulsion.

> If one asks how one's mind works, he notices areas where it is (perhaps incorrectly) understood, that is, where one recognizes rules. One sees other areas where he lacks rules. One could fill this in by postulating chance or random activity. But this too, by another route, exposes the self to the . . . indignity of remote control. We resolve this unpleasant form of M** by postulating a third part, embodying a will or spirit or conscious agent. But there is no structure in this part; one can say nothing meaningful about it, because whenever a regularity is observed, its representation is transferred to the deterministic rule region. The will model is thus not formed from a legitimate need for a place to store definite information about one's self; it has the singular character of being forced into the model, willy-nilly, by formal but essentially content-free ideas of what the model must contain.

One can quibble with the details, but the conceptual framework offers a whole new way of thinking about consciousness, by showing that introspection is mediated by models. There is no way for us to penetrate through them or shake them off, so we must simply live with any "distortion" they introduce. I put "distortion" in quotes because it's too strong a word. The concepts we use to describe our mental lives were developed over centuries by people who all shared the same kind of mental model. The distortions are built in. For instance, there is no independent notion of "free will" beyond what we observe by means of our self-models. We can't even say that free will is a dispensable illusion, because we have no way of getting rid of it and living to tell the tale. Minsky's insight is that to answer many questions about consciousness we should focus more on the models we use to answer the questions than on the questions themselves.

Unfortunately, in that short paper, and in his later book *The Society of Mind* (Minsky, 1986), Minsky throws off many interesting ideas, but refuses to go into the depth that many of them deserve. He has a lot to say about consciousness in passing, such as how Freudian phenomena might arise out of the "society" of subpersonal modules that he takes the human mind to be. But there is no solid proposal to argue for or against.

John McCarthy has written a lot on what he usually calls "self-awareness" (McCarthy, 1995b). However, his papers are mostly focused on robots' problem-solving capacities and how they would be enhanced by the ability to introspect. An important example is the ability of a robot to infer that it doesn't know something (such as whether the Pope is currently sitting or lying down). This may be self-awareness, but the word "awareness" here is used in a sense that is quite separate from the notion of phenomenal consciousness that is our concern here.

In (McCarthy, 1995a), he specifically addresses the issue of "zombies," philosophers' term for hypothetical beings who behave exactly as we do but do not experience anything. This paper is a reply to an article by Todd Moody (1994) on zombies. He lists some introspective capacities it would be good to give to a robot ("...Observing its goal structure and forming sentences about it .... Observing how it arrived at its current beliefs ...."). Then he concludes abruptly:

> Moody isn't consistent in his description of zombies. On page 1 they behave like humans. On page 3 they express puzzlement about human consciousness. Wouldn't a real Moody zombie behave as though it understood as much about consciousness as Moody does?

I tend to agree with McCarthy that the idea of a zombie is worthless, in spite of its initial plausibility. Quoting Moody:

> Given *any* functional [=, more or less, computational] description of cognition, as detailed and complete as one can imagine, it will still make sense to suppose that there could be insentient beings that exemplify that description. That is, it is possible that there could be a behaviourally indiscernible but insentient simulacrum of a human cognizer: a zombie.

The plausibility of this picture is that it does indeed seem that an intricate diagram of the hardware and software of a robot would leave consciousness out, just as with the computer-controlled heating system described in section 1. One could print it on rose-colored paper to indicate that the system was conscious, but the color of the paper would play no role in what it actually did. The problem is that in imagining a zombie one tends at first to forget that the zombie would say exactly the same things non-zombies say about their experiences. It would be very hard to convince a zombie that it lacked experience; which means, as far as I can see, that we might be zombies, at which point the whole idea collapses.

Almost everyone who thinks the idea is coherent sooner or later slips up the way Moody does: they let the zombie *figure out* that it is a zombie by noticing that it has no experience. By hypothesis, this is something zombies can't do. Moody's paper is remarkable only in how obvious the slip-up in it is.

> Consider, for example, the phenomenon of dreaming. Could there be a cognate concept in zombie-English? How might we explain dreaming to them? We could say that dreams are things that we experience while asleep, but the zombies would not be able to make sense[z] of this.[6]

Of course, zombies would talk about their dreams (or dreams[z]?) exactly as we do; consult the intricate system diagram to verify this.

McCarthy's three-sentence reply is just about what Moody's paper deserves. But meanwhile philosophers such as Chalmers (1996) have written weighty tomes based on the assumption that zombies make sense. McCarthy is not interested in refuting them.

Similarly, in (McCarthy, 1990b), McCarthy discusses when it is legitimate to ascribe mental properties to robots. In some ways his treatment is more formal than that of Dennett, which I discuss below. But he never builds on this theory to ask the key question: Is there more to your having a mental state than having that state ascribed to you?

3.3 Daniel Dennett

Daniel Dennett is not a researcher in artificial intelligence, but a philosopher of mind and essayist in cognitive science. Nonetheless, he is sympathetic to the AI project, and bases his philosophy on computational premises to a great degree. The models of mind that he has proposed can be considered to be sketches of a computational model, and therefore constitute one of the most ambitious and detailed proposals for how AI might account for consciousness.

Dennett's (1969) Ph.D. dissertation proposed a model for a conscious system. It contains the sort of block diagram that has since become a standard feature of the theories of psychologists such as Bernard Baars (1988, 1997), although the central working arena is designed to account for introspection more than for problem-solving ability.

[6] The "[z]" is used to flag zombie words whose meanings mustn't be confused with normal human concepts.

In later work, Dennett has not built upon this model, but, in a sense, has been rebuilding it from the ground up. The result has been a long series of papers and books, rich with insights about consciousness, free will, and intentionality. Their very richness makes it hard to extract a brisk theoretical statement, but I will try.

Dennett has one overriding methodological principle, to be distrustful of introspection. This position immediately puts him at odds with such philosophers as Nagel, Searle, and McGinn, for whom the "first person" point of view is the alpha and omega of consciousness. On his side Dennett has the many anecdotes and experimental data that show how wildly inaccurate introspection can be, but his view does leave him open to the charge that he is ruling out all the competitors to his theory from the start. From a computationalist's vantage point, this is all to the good. It's clear that any computationalist theory must eventually explain the mechanism of the first-person view in terms of "third person" components. The "third person" is that which you and I discuss, and therefore must be observable by you and me, and other interested parties, in the same way. In other words, "third-person data" is just another way of saying "scientific data." If there is to be a scientific explanation of the first person, it will surely seem more like an "explaining away" than a true explanation. An account of how yonder piece of meat or machinery is conscious will almost certainly invoke the idea of the machinery playing a trick on itself the result of which is for it to have a strong belief that it has a special first-person viewpoint.

One of Dennett's special skills is using vivid images to buttress his case. He invented the phrase "Cartesian Theater" to describe the hypothetical place in the brain where the self becomes aware of things. He observes that belief in the Cartesian Theater is deep-seated, and keeps popping up in philosophical and psychological writings, as well as in common-sense musings. We all know that there is a lot going on the brain that is preconscious or subconscious. What happens when a train of events becomes conscious? According to the view Dennett is ridiculing, to bring it to consciousness is to show it on the screen in the Cartesian Theater. When presented this way, the idea does seem silly, if for no other reason than that there is no plausible homunculus to put in the audience. What's interesting is how hard it is to shake this image. Just about all theorists of phenomenal consciousness at some point distinguish between "ordinary" and "conscious" events by making the latter be accessible to . . . what, exactly? The system as a whole? Its self-monitoring modules? One must tread very carefully to keep from describing the agent with special access as the good old transcendental self, sitting alone in the Cartesian Theater.

To demolish the Cartesian Theater, Dennett uses the tool of discovering or inventing situations in which belief in it leads to absurd conclusions. Many of these situations are experiments set up by psychology researchers. Most famous are the experiments by Libet (1985), whose object was to determine exactly when a decision to make a motion was made. What emerged from the experiments was that at the point where subjects *think* they have made the decision, the neural activity preparatory to the motion has already been in progress for hundreds of milliseconds. Trying to make sense of these results using the homuncular models lead to absurdities. (Perhaps the choice causes effects in the person's past?) But it is easy to explain them if you make a more inclusive picture of what's going on in a subject's brain. Libet and others tended to assume that giving a subject a button to push when the decision had been made provided a direct route to ... that pause again ... the subject's self, perhaps? Or perhaps the guy in the theater? Dennett points out that the neural apparatus required to push the button is part of the overall brain system. Up to a certain resolution, it makes sense to ask someone, "When did you decide to do X?" But it makes no sense to try to tease off a subsystem of the brain and ask it the same question, primarily because there is no subsystem that embodies the "will" of the whole system.

Having demolished most of the traditional model of consciousness, Dennett's next goal is to construct a new one, and here he becomes more controversial, and in places more obscure. A key component is human language. It is difficult to think about human consciousness without pondering the ability of a normal human adult to *say* what they are thinking. There are two possible views about why it should be the case that we can introspect so easily. One is that we evolved from animals that can introspect, so naturally when language evolved one of the topics it was used on was the contents of our introspections. The other is that language plays a more central role than that; without language, we wouldn't be conscious at all, at least full-bloodedly. Dennett's view is the second. He has little to say about animal consciousness, and what he does say is disparaging.

Language, for Dennett, is very important, but not because it is spoken by the homunculus in the Cartesian Theater. If you leave it out, who is speaking? Dennett's answer is certainly bold: In a sense, the language speaks itself. We take it for granted that speaking feels like it emanates from our "transcendental self," or, less politely, from the one-person audience in the Theater. Whether or not that view is correct now, it almost certainly was not correct when language began. In its original form, language was an information-transmission device used by apes whose consciousness, if similar to ours in any real respect, would be about the same as a chimpanzee's today. Messages expresssed linguistically would heard by one person, and for one reason or another be passed to others. Their chance of being passed would depend, very roughly, on how useful their recipients found them.

The same mechanism has been in operation ever since. Ideas (or simple patterns unworthy of the name "idea" — advertising jingles, for instance) tend to proliferate in proportion to how much they help those who adopt them, or in proportion to how well they tend to stifle competing ideas — not unlike what genes do. Dennett adopts Dawkins's (1976) term *meme* to denote a linguistic pattern conceived of in this way. One key meme is the idea of talking to oneself; when it first popped up, it meant literally talking out loud and listening to what was said. Although nowadays we tend to view talking to oneself as a possible symptom of insanity, we've forgotten that it gives our brains a whole new channel for its parts to communicate with each other. If an idea — a pattern of activity in the brain — can reach the linguistic apparatus, it gets translated into a new form, and, as it is heard, gets translated back into a somewhat different pattern than the one that started the chain of events. Creatures that start to behave this way start to think of themselves in a new light, as someone to talk to or listen to. Self modeling, according to Dennett (and Jaynes, 1976) starts as modeling this person we're talking to. There is nothing special about this kind of model; it is as crude as most of the models we make. But memes for self-modeling have been some of the most successful in the history (and prehistory) of humankind. To a great degree, they make us what we are by giving us a model of who we are that we then live up to. Every child must recapitulate the story Dennett tells, as it absorbs from its parents and peers all the ways to think of oneself, as a being with free will, sensations, and a still small voice inside.

The theory has one striking feature: it assumes that consciousness is based on language and not vice versa. For that matter, it tends to assume that for consciousness to come to be, there must be in place a substantial infrastructure of perceptual, motor, and intellectual skills. There may be some linguistic abilities that depend on consciousness, but the basic ability must exist *before* and *independent of* consciousness.

This conclusion may be fairly easy to accept for the more syntactic aspects of language, but it is contrary to the intuitions of many when it comes to semantics. Knowing what a sentence means requires knowing how the sentence relates to the world. If I am told "There is a lion on the other side of that bush," I have to understand that "that bush" refers to a particular object in view; I have to know how phrases like "other side of" work; and I have to understand what "a lion" means so that I have a grasp of just what I'm expecting to confront. Furthermore, it's hard to see how I could know what these words and phrases meant without knowing that I know what they mean.

Meditating in this way on how meaning works, the late-nineteenth-century philosopher Franz Brentano developed the notion of *intentionality*, the power mental representations seem to have of pointing to — "being about" — things outside of, and arbitrarily far from, the mind or brain containing those representations. The ability of someone to warn me about that lion depends on that person's sure-footed ability to reason about that animal over there, as well as on our shared knowledge about the species `Panthera leo`. Brentano, and many philosophers since, have argued that intentionality is at bottom a property *only* of mental representations. There seem to be many kinds of "aboutness" in the world; for instance, there are books about lions; but items like books can be about a topic only if they are created by humans using language and writing systems in order to capture thoughts about that topic. Books are said to have *derived* intentionality, whereas people have *original* or *intrinsic* intentionality.

Computers seem to be textbook cases of physical items whose intentionality, if any, is derived. If one sees a curve plotted on a computer's screen, the surest way to find out what it's about is to ask the person who used some program to create it. In fact, that's the *only* way. Digital computers are syntactic engines *par excellence.* Even if there is an interpretation to be placed on every step of a computation, this interpretation plays no role in what the computer does. Each step is produced purely by operations dependent on the formal structure of its inputs and prior state at that step. If you use TurboTax to compute your income taxes, then the numbers being manipulated represent real-world quantities, and the number you get at the end represents what you actually do owe to the tax authorities. Nonetheless, TurboTax is just applying formulas to the numbers. It "has no idea" what they mean.

This intuition is what Dennett wants to defeat, as should every other researcher who expects a theory of consciousness based on AI. There's really no alternative. If you believe that people are capable of original intentionality and computers aren't, then you must believe that something will be missing from any computer program that tries to simulate humans. That means that human consciousness is fundamentally different from machine consciousness, which means that a theory of consciousness based on AI is radically incomplete.

Dennett's approach to the required demolition job on intrinsic intentionality is to focus on the prelinguistic, nonintrospective case. In a way, this is changing the subject fairly radically. In the introspective set-up, we are talking about elements or aspects of the mind that we are routinely acquainted with, such as words and images. In the nonintrospective case, it's not clear that those elements or aspects are present at all. What's left to talk about if we're not talking about words, "images," or "thoughts"? We'll have to shift to talking about neurons, chips, firing rates, bits, pointers, and other "subpersonal" entities and events. It's not clear at all whether these things are even capable of exhibiting intentionality. Nonetheless, showing that they are is a key tactic in Dennett's attack on the problem of consciousness. (See especially Appendix A of Dennett, 1991b.) If we can define what it is for subpersonal entities to be intentional, we can then build on that notion and recover the phenomenal entities we (thought we) started with. "Original" intentionality will turn out to be a secondary consequence of what I will call *impersonal intentionality.*

Dennett's approach to the problem is to call attention to what he calls the *intentional stance,* a way of looking at systems in which we impute beliefs and goals to them simply because there's no better way to explain what they're doing. For example, if you're observing a good chess program in action, and its opponent has left himself vulnerable to an obvious attack, then one feels confident that the program will embark on that attack. This confidence is not based on any detailed knowledge of the program's actual code. Even someone who knows the program well won't bother trying to do a tedious simulation to make a prediction that the attack will occur, but will base their prediction on the fact that the program almost never misses an opportunity of that kind. If you refuse to treat the program as though it had goals, you will be able to say very little about how it works.

The intentional stance applies to the innards of the program as well. If a data structure is used by the program to make decisions about some situation or object S, and the decisions it makes are well explained by assuming that one state of the data structure means that P is true of S, and that another means P', then those states *do* mean P and P'.

It is perhaps unfortunate that Dennett has chosen to express his theory this way, because it is easy to take him as saying that all intentionality is observer-relative. This would be almost as bad as maintaining a distinction between original and derived intentionality, because it would make it hard to see how the process of intentionality attribution could ever get started. Presumably my intuition that I am an intentional system is indubitable, but what could it be based on? It seems absurd to think that this opinion is based on what others tell me, but it seems equally absurd that I could be my own observer. Presumably to be an observer you have to be an intentional system (at least, if your observations are to be *about* anything). Can I bootstrap my way into intentionality somehow? If so, how do I tell the successful bootstrappers from the unsuccessful ones? A computer program with an infinite loop, endlessly printing, "I am an intentional system because I predict, by taking the intentional stance, that I will continue to print this sentence out," would not actually be claiming anything, let alone something true.

Of course, Dennett does not mean for intentionality to be observer-relative, even though many readers think he does. (To take an example at random from the Internet, the on-line *Philosopher's Magazine,* in their "Philosopher of the Month" column in April, 2003 (Douglas & Saunders, 2003), say "Dennett suggests that intentionality is not so much an intrinsic feature of agents, rather, it is more a way of looking at agents.") Dennett has defended himself from this misinterpretation more than once (Dennett, 1991a). I will come back to this issue in my attempt at a synthesis in section 4.

3.4 Perlis, Sloman

The researchers in this section, although they work on hard-headed problems in artificial intelligence, do take philosophical problems seriously, and have contributed substantial ideas to the development of the computational model of consciousness.

Donald Perlis's papers build a case that consciousness is ultimately based on self-consciousness, but I believe he is using the phrase "self-consciousness" in a misleading and unnecessary way. Let's start with his paper (Perlis, 1994), which I think lays out a very important idea. He asks, Why do we need a dichotomy between appearance and reality? The answer is, Because they could disagree, i.e., because I could be wrong about what I think I perceive. For an organism to be able to reason explicitly about this difference, it must be able to represent both X (an object in the world) and quote-X, the representation of X in the organism itself. The latter is the "symbol," the former the "symboled." To my mind the most important consequence of this observation is that it must be possible for an information processing system to get two kinds of information out of its X-recognizer: signals meaning "there's an X," and signals meaning "there's a signal meaning 'there's an X.' "

Perlis takes a somewhat different tack. He believes there can be no notion of appearance without the notion of appearance *to* someone. So the self-model can't get started without some prior notion of self to model.

> When we are conscious of X, we are also conscious of X in relation to ourselves: it is here, or there, or seen from a certain angle, or thought about this way and then that. Indeed, without a self model, it is not clear to me intuitively what it means to see or feel something: it seems to me that a point of view is needed, a place from which the scene is viewed or felt, defining the place occupied by the viewer. Without

> something along these lines, I think that a 'neuronal box' would indeed 'confuse' symbol and symboled: to it there is no external reality, it has no way to 'think' (consider alternatives) at all. Thus I disagree [with Crick] that self-consciousness is a special case of consciousness: I suspect that it is the most basic form of all.

Perlis continues to elaborate this idea in later publications. For example, "*Consciousness is the function or process that allows a system to distinguish itself from the rest of the world*.. . . To feel pain or have a vivid experience requires a self" (Perlis, 1997) (italics in original). I have trouble following his arguments, which often depend on thought experiments such as imagining cases where one is conscious but not *of* anything, or of as little as possible. The problem is that introspective thought experiments are just not a very accurate tool. One may perhaps conclude that Perlis, although housed in a Computer Science department, is not a thoroughgoing computationalist at all. As he says, "I conjecture that we may find in the brain special amazing structures that facilitate true self-referential processes, and constitute a primitive, bare or ur-awareness, an 'I'. I will call this the *amazing-structures-and-processes paradigm*" (Perlis, 1997) (italics in original). It's not clear how amazing the "amazing" structures will be, but perhaps they won't be computational.

Aaron Sloman has written prolifically about philosophy and computation, although his interests range far beyond our topic here. In fact, although he has been interested in conscious control, both philosophically and as a strategy for organizing complex software, he has tended to shy away from the topic of phenomenal consciousness. His book *The Computer Revolution in Philosophy* (Sloman, 1978) has almost nothing to say about the subject, and in many other writings the main point he has to make is that the concept of consciousness covers a lot of different processes, which should be sorted out before hard questions can be answered. However, in a few of his papers he has confronted the issue of qualia, notably (Sloman & Chrisley, 2003). I think the following is exactly right:

> Now suppose that an agent A . . . uses a self-organising process to develop concepts for categorising its own internal virtual machine states as sensed by internal monitors. . . . If such a concept C is applied by A to one of its internal states, then the only way C can have meaning for A is in relation to the set of concepts of which it is a member, which in turn derives only from the history of the self-organising process in A. These concepts have what (Campbell, 1994) refers to as 'causal indexicality'. This can be contrasted with what happens when A interacts with other agents in such a way as to develop a common language for referring to features of external objects. Thus A could use 'red' either as expressing a private, causally indexical, concept referring to features of A's own virtual-machine states, or as expressing a shared concept referring to a visible property of the surfaces of objects. This means that if two agents A and B have each developed concepts in this way, then if A uses its causally indexical concept Ca, to think the thought 'I am having experience Ca',

> and B uses its causally indexical concept Cb, to think the thought ‘I am having experience Cb’ the two thoughts are intrinsically private and incommunicable, even if A and B actually have exactly the same architecture and have had identical histories leading to the formation of structurally identical sets of concepts. A can wonder: ‘Does B have an experience described by a concept related to B as my concept Ca is related to me?’ But A cannot wonder ‘Does B have experiences of type Ca’, for it makes no sense for the concept Ca to be applied outside the context for which it was developed, namely one in which A’s internal sensors classify internal states. They cannot classify states of B.

This idea suggests that the point I casually assumed at the beginning of this paper, that two people might wonder if they experienced the same thing when they ate tacos, is actually incoherent. Our feeling that the meaning is clear is due to the twist our self-models give to introspections of the kind Sloman and Chrisley are talking about. The internal representation of the quale of redness is purely local to A’s brain, but the self-model says quite the opposite, that objects with the color are recognizable by A because they have that quale. The quale is made into an objective entity that might attach itself to other experiences, such as my encounters with blue things, or B’s experiences of red things.

3.5 Brian Cantwell Smith

The last body of research to be examined in this survey is that of Brian Cantwell Smith. It’s hard to dispute that he is a computationalist, but he is also an antireductionist, which places him in a rather unique category. Although it is clear in reading his work that he considers consciousness to be a crucial topic, he has been working up to it very carefully. His early work (Smith, 1984) was on “reflection” in programming languages, that is, how and why a program written in a language could have access to information about its own subroutines and data structures. One might conjecture that reflection might play a key role in a system’s maintaining a self-model and thereby being conscious. But since that early work Smith has moved steadily away from straightforward computational topics and toward foundational philosophical ones. Each of his papers seems to take tinier steps from first principles than the ones that have gone before, so as to presuppose as little as humanly possible. Nonetheless, they often express remarkable insight. His paper (Smith, 2002) on the “Foundations of Computing” is a gem. (I also recommend (Sloman, 2002), from the same collection (Scheutz, 2002).)

One thing both Smith and Sloman argue is that Turing machines are misleading as ideal vehicles for computationalism, which is a point often missed by philosophers. For example, Wilkes (1990) says “...computers (as distinct from robots) produce at best only linguistic and exclusively ‘cognitive’ — programmable — ‘behaviour’: the emphasis is on internal psychological processes, the cognitive ‘inner’ rather than on action, emotion, motivation, and sensory experience.” Perhaps I’ve misunderstood him, but it’s very hard to see how this can be true, given that all interesting robots are controlled by digital computers. Furthermore, when computers and software are studied isolated from their physical environments, it's often for purely tactical reasons (from budget or personnel limitations, or to avoid endangering bystanders). If we go all the way back to Winograd's (1972) SHRDLU system, we find a simulated robot playing the role of conversationalist, not because Winograd thought real robots were irrelevant, but precisely because he was thinking of a long-term project in which an actual robot would be used.

As Smith (2002) says,

> In one way or another, no matter what construal [of formality] they pledge allegiance to, just about everyone thinks that computers are formal. . . . But since the outset, I have not believed that this is necessarily so.. . . Rather, what computers are . . . is neither more nor less than the *full-fledged social construction and development of intentional artifacts.* (italics in original)

The point he is trying to make (and it can be hard to find a succinct quote in Smith's papers) is that computers are *always* connected to the world, whether they are robots or not, and therefore the meaning their symbols possess is more determined by those connections than by what a formal theory might say they mean. One might want to rule that the transducers that connect them to the world are noncomputational (cf. (Harnad, 1990)), but there is no principled way to draw a boundary between the two parts, because ultimately a computer is physical parts banging against other physical parts. As Sloman puts it,

> . . . The view of computers as somehow essentially a form of Turing machine . . . is simply mistaken. . . . [The] mathematical notion of computation . . . is not the primary motivation for the construction or use of computers, nor is it particularly helpful in understanding how computers work or how to use them (Sloman, 2002).

The point Smith makes in the paper cited above is elaborated into an entire book, *On the Origin of Objects* (Smith, 1995). The problem the book addresses is the basic ontology of physical objects. The problem is urgent, according to Smith, because the basic concept of intentionality is that a symbol S stands for an object X; but we have no prior concept of what objects or symbols are. A geologist might see a glacier on a mountain, but is there some objective reason why the glacier is an object (and the group of stones suspended in it is not)? Smith believes that all object categories are to some extent carved out by subjects, i.e., by information-processing systems like us (and maybe someday by robots as well).

The problem with this point of view is that it is hard to bootstrap oneself out of what Smith calls the Criterion of Ultimate Concreteness: "No naturalistically palatable theory of intentionality — of mind, computation, semantics, ontology, objectivity — can presume the identify or existence of any individual object whatsoever" (p. 184). He tries valiantly to derive subjects and objects from prior . . . umm . . . "entities" called s-regions and o-regions, but it is hard to see how he succeeds. In spite of its length, 420 pages, the book claims to arrive at no more than a starting point for a complete rethinking of physics, metaphysics, and everything else.

Most people will have a hard time following Smith's inquiry, not least because few people agree on his opening premise, that everyday ontology is broken and needs to be fixed. I actually do agree with that, but I think the problem is much worse than Smith does. Unlike him, I am reductionist enough to believe that physics is the science of "all there is"; so how do objects emerge from a primordial superposition of wave functions? Fortunately, I think this is a problem for everyone, and has nothing to do with the problem of intentionality.[7] If computationalists are willing to grant that there's a glacier over there, anyone should be willing to consider the computational theory of how systems refer to glaciers.

4 A Synthetic Summary

In spite of the diffidence of most AI researchers on this topic, I believe that there is a dominant position on phenomenal consciousness among computationalists, "dominant" in the sense that among the small population of those who are willing to take a clear position, this is more or less the position they take. In this section I will try to sketch that postion, pointing out the similarities and differences from the positions sketched in section 3.

The idea in a nutshell is that phenomenal consciousness is the property a computational system X has if X models itself as experiencing things. To understand it, I need to explain

1. What a computational system is.
2. How such a system can exhibit intentionality.
3. That to be conscious is to model oneself as having experiences.

4.1 The Notion of Computational System

Before we computationalists can really get started, we run into the objection that the word "computer" doesn't denote the right kind of thing to play an explanatory role in a theory of any natural phenomenon. A computer, so the objection goes, is an object that people[8] *use to compute things.* Without people to assign meanings to its inputs and outputs, a computer is just an overly complex electronic kaleidoscope, generating a lot of pseudo-random patterns. We may interpret the output of a computer as a prediction about tomorrow's weather, but there's no other sense in which the computer is predicting anything. A chess computer outputs

[7] Even more fortunate, perhaps, is the fact that few will grant that foundational ontology is a problem in the first place. Those who think elementary particles invented us, rather than vice versa, are in the minority.

[8] Or intelligent aliens, but this is an irrelevant variation on the theme.

a syntactically legal expression that we can take to be its next move, but the computer doesn't actually intend to make that move. It doesn't intend *anything*. It doesn't care whether the move is actually made. Even if it's displaying the move on a screen, or using a robot arm to pick up a piece and move it, these outputs are just meaningless pixel values or drive-motor torques until *people* supply the meaning.

In my opinion, the apparent difficulty of supplying an objective definition of syntax and especially semantics is the most serious objection to the computational theory of psychology, and in particular to a computational explanation of phenomenal consciousness. To overcome it, we need to come up with a theory of computation (and eventually semantics) that is observer-independent.

There are two prongs to this attack, one syntactic, the other semantic. The syntactic prong is the claim that even the symbols we attribute to computers are observer-relative. We point to a register in the computer's memory, and claim that it contains a number. The critic then says that the mapping of states that cause this state to encode "55,000" is entirely arbitrary; there are an infinite number of ways of interpreting the state of the register, none of which is the "real" one in any sense; all we can talk about is the *intended* one. A notorious example of John Searle's exemplifies this kind of attack; he claims in (Searle, 1992) that the wall of his office could be considered to be a computer under the right encoding of its states.

The semantic prong is the observation, discussed in sections 3.3 and 3.5, that even after we've agreed that the register state encodes "55,000," there is no objective sense in which this figure stands for "Jeanne D'Eau's 2003 income in euros." If Jeanne D'Eau is using the EuroTax software package to compute her income tax, then such semantic statements are nothing but a convention adopted by her and the people that wrote EuroTax. In other words, the only intentionality exhibited by the program is derived intentionality.

To avoid these objections, we have to be careful about how we state our claims. I have space for only a cursory overview here; see (McDermott, 2001) for a more detailed treatment. First, the idea of "computer" is prior to the idea of "symbol." A *basic computer* is any physical system whose subsequent states are predictable given its prior states. By "state" I mean "partial state," so that the system can be in more than one state at a time. An *encoding* is a mapping from partial physical states to some syntactic domain (e.g., numerals). To view a system as a computer, we need two encodings, one for inputs, one for outputs. It computes $f(x)$ with respect to a pair $\langle I, O \rangle$ of encodings if and only if putting it into the partial state encoding x under I causes it to go into a partial state encoding $f(x)$ under O.

A *memory element* under an encoding E is a physical system that, when placed into a state s such that $E(s) = x$, tends to remain in the set of states $\{s : E(s) = x\}$ for a while.

A *computer* is then a group of basic computers and memory elements viewed under a consistent encoding scheme, meaning merely that if changes of component 1's state cause component 2's state to change, then the encoding of 1's outputs is the same as the encoding of 2's inputs. *Symbol sites* then appear as alternative possible stable regions of state space, and *symbol tokens* as chains of symbol sites such that the occupier of a site is caused by the presence of the occupier of its predecessor site. Space does not allow me to discuss all the details here, but the point is clear: the notions of *computer* and *symbol* are not observer-relative. Of course, they are *encoding-relative*, but then velocity is "reference-frame-relative." The encoding is purely syntactic, or even pre-syntactic, since we have said nothing about what syntax an encoded value has, if any. We could go on to say more about syntax, but one has the feeling that the whole problem is a practical joke played by philosophers on naive AI researchers. ("Let's see how much time we can get them to waste defining 'computer' for us, until they catch on.") I direct you to (McDermott, 2001) for more of my theory of syntax. The important issue is semantics, to which we now turn.

One last remark: The definitions above are not intended to distinguish digital from analogue computers, or serial from parallel ones. They are broad enough to include anything anyone might ever construe as a computational system. In particular, they allow neural nets (Rumelhart et al., 1986), natural and artificial, to count as computers. Many observers of AI (Churchland, 1986; Churchland, 1988; Wilkes, 1990) believe that there is an unbridgeable chasm between some classical, digital, traditional AI and a revolutionary, analogue, connectionist alternative. The former is the realm of von Neumann machines, the latter the realm of artificial neural networks, "massively parallel" networks of simple processors (meant to mimic neurons), which can be trained to learn different categories of sensory data (Rumelhart et al., 1986). The "chasm" between the two is less observable in practice than you might infer from the literature. AI researchers are omnivorous consumers of algorithmic techniques, and think of neural nets as one of them — entirely properly, in my opinion. I will return to this subject in section 5.3.

4.2 Intentionality of Computational Systems

I described above Dennett's idea of the "intentional stance," in which an observer explains a system's behavior by invoking intentional categories such as beliefs and goals. Dennett is completely correct that there is such a stance. The problem is that we sometimes adopt it inappropriately. People used to think thunderstorms were out to get them, and a sign on my wife's printer says, "Warning! This machine is subject to breakdown during periods of critical need." What could it possibly mean to say that a machine demonstrates *real* intentionality when it is so easy to indulge in a mistaken or merely metaphorical "intentional stance"?

Let's consider an example. Suppose someone has a cat that shows up in the kitchen at the time it is usually fed, meowing and behaving in other ways that tend to attract the attention of the people who usually feed it. Contrast that with the case of a robot that, whenever its battery is low, moves along a black trail painted on the floor that leads to the place where it gets recharged, and, when it is over a large black cross that has been painted at the end of the trail, emits a series of beeps that tend to attract the attention of the people who usually recharge it. Some people might refuse to attribute intentionality to either the cat or the robot, and treat comments such as, "It's trying to get to the kitchen [or recharging area]," or "It wants someone to feed [or recharge] it," as purely metaphorical. They might take this position, or argue that it's tenable, on the grounds that we have no reason to suppose that either the cat or the robot has mental states, and hence nothing with the kind of "intrinsic aboutness" that people exhibit. High catologists[9] are sure cats do have mental states, but the skeptic will view this as just another example of someone falling into the metaphorical pit of "as-if" intentionality.

I believe, though, that even hard-headed low catologists think the cat is truly intentional, albeit in the impersonal way discussed in section 3.3. They would argue that if you could open up its brain you would find neural structures that "referred to" the kitchen or the path to it, in the sense that those structures became active in ways appropriate to the cat's needs: they were involved in steering the cat to the kitchen and stopping it when it got there. A similar account would tie the meowing behavior to the event of getting food, mediated by some neural states. We would then feel justified in saying that some of the neural states and structures *denoted* the kitchen, or the event of being fed.

The question is, Are the ascriptions of impersonal intentionality so derived arbitrary, or objectively true? It's difficult to take either choice. It feels silly saying that something is arbitrary if it takes considerable effort to figure it out, and if one is confident that if someone else independently undertook the same project they would reach essentially the same result. But it also feels odd to say that something is objectively true if it is *inherently* invisible. Nowhere in the cat will you find labels that say "This means X," nor little threads that tie neural structures to objects in the world. One might want to say that the cat is an intentional system because there was evolutionary pressure in favor of creatures whose innards were tied via "virtual threads" to their surroundings. I don't like dragging evolution in because it's more of a question stopper than a question answerer. I prefer the conclusion that reluctance to classify intentionality as objectively real simply reveals an overly narrow conception of objective reality.

A couple of analogies will help.

1. Code breaking: A code breaker is sure they have cracked a code when the message turns into meaningful natural-language text. That's because there are an enormous number of possible messages, and an enormous number of possible ciphers, out of which there is (almost certainly) only one combination of natural-language text and simple cipher that produces the encrypted message.

   Unfortunately for this example, it involves interpreting the actions of people. So even if there is no observer-relativity from the cryptanalyst's point of view, the intentionality in a message is "derived" according to skeptics about the possible authentic intentionality of physical systems.

2. Geology: A geologist strives to find the best explanation for how various columns and strata of rock managed to place themselves in the positions they are found in. A good

[9] By analogy with Christology in Christian theology, which ranges from high to low depending on how superhuman one believes Jesus to be.

> explanation is a series of not-improbable events that would have transformed a plausible initial configuration of rocks into what we see today.
>
> In this case, there is no observer-relativity, because there was an actual sequence of events that led to the current rock configuration. If two geologists have a profound disagreement about the history of a rock formation, they can't both be right (as they might be if disagreeing about the beauty of a mountain range). Our normal expectation is that any two geologists will tend to agree on at least the broad outline of an explanation of a rock formation; and that as more data are gathered the areas of agreement will grow.

These examples are cases where, even though internal harmoniousness is how we judge explanations, what we get in the end is an explanation that is true, *independent of the harmoniousness*. All we need to do is allow for this to be true even though, in the case of intentionality, even a time machine or mind reader would not give us an independent source of evidence. To help us accept this possibility, consider the fact that geologists can never actually get the entire story right. What they are looking at is a huge structure of rock with a detailed microhistory that ultimately accounts for the position of every pebble. What they produce in the end is a coarse-grained history that talks only about large intrusions, sedimentary layers, and such. Nonetheless we say that it is objectively true, even though the objects it speaks of don't even exist unless the account is true. It explains how a particular "intrusion" got to be there, but if geological theory isn't more or less correct, there might not be such a thing as an intrusion; the objects might be parsed in a totally different way.

If processes and structures inside a cat's brain exhibit objectively real impersonal intentionality, then it's hard not to accept the same conclusion about the robot trying to get recharged. It might not navigate the way the cat does — for instance, it might have no notion of a place it's going to, as opposed to the path that gets it there — but we see the same fit with its environment among the symbol structures in its hardware or data. In the case of the robot the hardware and software were designed, and so we have the extra option of asking the designers what the entities inside the robot were *supposed* to denote. But it will often happen that there is conflict between what the designers intended and what actually occurs, and *what actually occurs wins*. The designers don't get to say, "This boolean variable means that the robot is going through a door" unless the variable's being **true** tends to occur if and only if the robot is between two door jambs. If the variable is correlated with something else instead, then *that's* what it actually means. It's appropriate to describe what the roboticists are doing as debugging the robot so that its actual intentionality matches their intent. The alternative would be to describe the robot as "deranged" in the sense that it continuously acts in ways that are bizarre given what its data structures mean.

Two other remarks are in order: What the symbols in a system mean is dependent on the system's environment. If a cat is moved to a house that is so similar to the one it's familiar with that the cat is fooled, then the structures inside it that used to refer to the kitchen of house 1 now refer to the kitchen of house 2. And so forth; and there will of course be cases in which the denotation of a symbol breaks down, leaving no coherent story about what it denotes, just as in the geological case an event of a type unknown to geology, but large enough to cause large-scale effects, will go unhypothesized, and some parts of geologists' attempts to make sense of what they see will be too incoherent to be true or false, or to even to refer to anything.

The other remark is that it might be the case that the sheer size of the symbolic systems inside people's heads might make the impersonal intentionality story irrelevant. We don't, of course, know much about the symbol systems used by human brains, whether there is a "language of thought" (Fodor, 1975) or some sort of connectionist soup, but clearly we can have beliefs orders of magnitude more complex than those of a cat or a robot (year-2006 model). If you walk to work, but at the end of the day absent-mindedly head for the parking lot to retrieve your car, what you will believe once you get there has the content "My car is not here." Does this belief correspond to a symbol structure in the brain whose pieces include symbol tokens for "my car," "here," and "not"? We don't know. But if anything like that picture is accurate, then assigning a meaning to symbols such as "not" is considerably more difficult than assigning a meaning to the symbols a cat or robot might use to denote "the kitchen." Nonetheless, the same basic story can still be told: that the symbols mean what the most harmonious interpretation says they mean. This story allows us to assign arbitrarily abstract meanings to symbols like "not"; the price we pay is that for now all we have is an IOU for a holistic theory of the meanings inside our heads.

4.3 Modelling Oneself as Conscious

I have spent a lot of time discussing intentionality because once we can establish the concept of an impersonal level of meaning in brains and computers, we can introduce the idea of a *self-model,* a device that a robot or a person can use to answer questions about how it interacts with the world. This idea was introduced by Minsky almost forty years ago (Minsky, 1968a), and has since been explored by many others, including Sloman (Sloman & Chrisley, 2003), McDermott (2001), and Dennett (1991b). As I mentioned above, Dennett mixes this idea with the concept of meme, but self-models don't need to be made out of memes.

We start with Minsky's observation that complex organisms use models of their environments in order to predict what will happen and decide how to act. In the case of humans, model making is taken for granted by psychologists (Johnson-Laird, 1983); no one really knows what other animals' capacities for using mental models are. A *mental model* is some sort of internal representation of part of the organism's surroundings that can be inspected, or even "run" in some way, so that features of the model can then be transformed back into inferred or predicted features of the world. For example, suppose you're planning to go grocery shopping, and the skies are threatening rain, and you're trying to decide whether to take an umbrella. You enumerate the situations where the umbrella might be useful, and think about whether on balance it will be useful enough to justify having to keep track of it. One such situation is the time when you emerge from the store with a cartload of groceries to put in the car. Will the umbrella keep you or your groceries dry?[10]

This definition is general (and vague) enough to cover noncomputational models, but the computationalist framework provides an obvious and attractive approach to theorizing about mental models. In this framework, a model is an internal computer set up to simulate something. The organism initializes it, lets it run for a while, reads off its state, and interprets the state as a set of inferences that then guide behavior. In the umbrella example, one might imagine a physical simulation, at some level of resolution, of a person pushing a cart and holding an umbrella while rain falls.

A mental model used by an agent A to decide what to do must include A itself, simply because any situation A finds itself in will have A as one of its participants. If I am on a sinking ship, and trying to pick a lifeboat to jump into, predicting the number of people on the lifeboat must not omit the "+ 1" required to include me. This seemingly minor principle has far-reaching consequences, because many of A's beliefs about itself will stem from the way its internal surrogates participate in mental models. We will call the beliefs about a particular surrogate a *self-model*, but usually for simplicity I will refer to *the* self-model, as if all those beliefs are pulled together into a single "database." Let me state up front that the way things really work is likely to be much more complex and messy. Let me also declare that the self-model is *not* a Cartesian point of transcendence where the self can gaze at itself. It is a resource accessible to the brain at various points for several different purposes.

We can distinguish between *exterior* and *interior* self-models. The former refer to the agent considered as a physical object, something with mass that might sink a lifeboat. The latter refers to the agent considered as an information-processing system. To be concrete, let's look at a self-model that arises in connection with the use of *anytime algorithms* to solve *time-dependent planning problems* (Boddy & Dean, 1989). An anytime algorithm is one that can be thought of as an asynchronous process that starts with a rough approximation to the desired answer and gradually improves it; it can be stopped at any time and the quality of the result it returns depends on how much run time it was given. We can apply this idea to planning robot behavior, in situations where the objective is to minimize the total time required to solve the problem, which is equal to

$$\text{time } (t_P) \text{ to find a plan } P + \text{time } (t_E(P)) \text{ to execute } P$$

If the planner is an anytime algorithm, then the quality of the plan it returns improves with t_P. We write $P(t_P)$ to indicate that the plan found is a function of the time allotted to finding it. Because quality is execution time, we can refine that statement and say that $t_E(P(t_P))$ decreases as t_P increases. Therefore, in order to optimize

$$t_P + t_E(P(t_P))$$

[10] For some readers this example will elicit fairly detailed visual images of shopping carts and umbrellas, and for those readers it's plausible that the images are part of the mental-model machinery. But even people without much visual imagery can still have mental models, and might still use them to reason about grocery shopping.

we must find the smallest t_P such that the time gained by planning Δt longer than that would probably improve t_E by less than Δt. The only way to find that optimal t_P is to have an approximate model of how fast $t_E(P(t_P))$ changes as a function of t_P. Such a model would no doubt reflect the law of diminishing returns, so that finding the optimal t_P is an easy one-dimensional optimization problem. The important point for us is that this model is a model of the planning component of the robot, and so counts as an interior self-model.

Let me make sure my point is clear: interior self-models are no big deal. Any algorithm that outputs an estimate of something plus an error range incorporates one. The mere presence of a self-model does not provide us some kind of mystical reflection zone where we can make consciousness pop out as an "emergent" phenomenon. This point is often misunderstood by critics of AI (Rey, 1997; Block, 1997) who attribute to computationalists the idea that consciousness is nothing but the ability to model oneself. In so doing, they tend to muddy the water further by saying that computationalists confuse consciousness with self-consciousness. I hope in what follows I can make these waters a bit clearer.

Today's information-processing systems are not very smart. They tend to work in narrow domains, and outperform humans only in areas, such as chess and numerical computation, where clear formal ground rules are laid out in advance. A robot that can walk into a room, spy a chessboard, and ask if anyone wants to play is still far in the future. This state of affairs raises a huge obstacle for those who believe that consciousness is built on top of intelligence rather than vice versa, that obstacle being that everything we say is hypothetical. It's easy to counter the computationalist argument. Just say, "I think you're wrong about intelligence preceding consciousness, but even if you're right I doubt that computers will ever reach the level of intelligence required."

To which I reply, Okay. But let's suppose they do reach that level. We avoid begging any questions by using my hypothetical chess-playing robot as a concrete example. We can imagine it being able to locomote, see chessboards, and engage in simple conversations. ("Want to play?" "Later." "I'll be back.") We start by assuming that it is not conscious, and then think about what it would gain by having interior self-models of a certain class. The starting assumption, that it isn't conscious, should be uncontroversial.

One thing such a robot might need is a way to handle perceptual errors. Suppose that it has a subroutine for recognizing chessboards and chessmen.[11] For serious play only Staunton chess pieces are allowed, but you can buy a chessboard with pieces of almost any shape; I have no doubt that Disney sells a set with Mickey and Minnie Mouse as king and queen. Our robot, we suppose, can correct for scale, lighting, and other variations of the appearance of Staunton pieces, but just can't "parse" other kinds of pieces. It could also be fooled by objects that only appeared to be Staunton chess pieces.

[11] I have two reasons for positing a chessboard-recognition subroutine instead of a general-purpose vision system that recognizes chessboards and chess pieces in terms of more "primitive" elements: (1) Many roboticists prefer to work with specialized perceptual systems; and (2) the qualia-like entities we will predict will be different in content from human qualia, which reduces the chances of jumping to conclusions about them.

Now suppose that the robot contained some modules for improving its performance. It might be difficult to calibrate the perceptual systems of our chess-playing robots at the factory, especially since different owners will use them in different situations. So we suppose that after a perceptual failure a module we will call the *perception tuner* will try to diagnose the problem and change the parameters of the perceptual system to avoid it in the future.

The perception tuner must have access to the inputs and outputs of the chess recognition system, and, of course, access to parameters that it can change in order to improve the system's performance. It must have a self-model that tells it how to change the parameters to reduce the likelihood of errors. (The "back propagation" algorithm used in neural nets (Rumelhart et al., 1986) is an example.) What I want to call attention to is that the perception tuner interprets the outputs of the perceptual system in a rather different way from the decision-making system. The decision-making system interprets them (to oversimplify) as being about the environment; the tuning system interprets them as being about the perceptual system. For the decision maker the output "Pawn at $\langle x, y, z\rangle$" means that there is a pawn at a certain place. For the tuner, it means that the perceptual system *says* there is a pawn, in other words, that there *appears to be* a pawn.

Here is where the computationalist analysis of intentionality steps in. We don't need to believe that either the decision maker or the tuner literally "thinks" that a symbol structure at a certain point means a particular thing. The symbol structure S means X if there is a harmonious overall interpretation of the states of the robot in which S means X.

The perceptual-tuner scenario suggests that we can distinguish two sorts of access to a subsystem: normal access and introspective access. The former refers to the flow of information that the subsystem extracts from the world (Dretske, 1981). The latter refers to the flow of information it produces *about* the normal flow.[12] For our robot, normal access gives it information about chess pieces; introspective access gives it information about . . . what, exactly? A datum produced by the tuner would consist of a designator of some part of the perceptual field that was misinterpreted, plus information about how it was interpreted and how it should have been. We can think of this as being information about "appearance" vs. "reality."

The next step in our story is to suppose that our robot has "episodic" memories, that is, memories of particular events that occurred to it. (Psychologists draw distinctions between these memories and other kinds, such as learned skills (e.g.,the memory of how to ride a bicycle) and abstract knowledge (e.g., the memory that France is next to Germany), sometimes called *semantic memory.*) We take episodic memory for granted, but presumably flatworms do without it; there must be a reason why it evolved in some primates. One possibility is that it's a means to keep track of events whose significance is initially unknown. If something bad or good happens to an organism, it might want to retrieve past occasions when something similar happened and

[12] Of course, what we'd like to be able to say here is that normal access is the access it was designed to support, and for most purposes that's what we will say, even when evolution is the "designer." But such basic concepts can't depend on historical events arbitrarily far in the past.

try to see a pattern. It's hard to say why the expense of maintaining a complex "database" would be paid back in terms of reproductive success, especially given how wrongheaded people can be about explaining patterns of events. But perhaps all that is required is enough paranoia to avoid too many false negatives in predicting catastrophes.

The final step is to suppose that the robot can ask fairly general questions about the operation of its perceptual and decision-making systems. Actually, this ability is closely tied to the ability to store episodic memories. To remember something one must have a notation to express it. Remembering a motor skill might require storing a few dozen numerical parameters (e.g., weights in neural networks, plus some sequencing information). If this is correct, then, as argued above, learning a skill means nudging these parameters toward optimal values. Because this notation is so lean, it won't support recording the episodes during which skill was enhanced. You may remember your golf lessons, but those memories are independent of the "memories," encoded as numerical parameters, that manifest themselves as an improved putt. If you try to think of a notation in which to record an arbitrary episode, it's like trying to think of a formal notation to capture the content of a Tolstoy novel. It's not even clear what it would mean to record an episode. How much detail would there be? Would it always have to be from the point of view of the creature that recorded it? Such questions get us quickly into the realm of Knowledge Representation, and the Language of Thought (Fodor, 1975). For that matter, we are quickly led to the topic of ordinary human language, because the ability to recall an episode seems closely related to the ability to tell about it, and to ask about it. We are far from understanding how language, knowledge representation, and episodic memory work, but it seems clear that the mechanisms are tightly connected, and all have to do with what sorts of questions the self-model can answer. This clump of mysteries accounts for why Dennett's (1991b) meme-based theory is so attractive. He makes a fairly concrete proposal that language came first and that the evolution of the self-model was driven by the evolution of language.

Having waved our hands a bit, we can get back to discussing the ability of humans, and presumably other intelligent creatures, to ask questions about how they work. We will just assume that these questions are asked using an internal notation reminiscent of human language, and then answered using a Minskyesque self-model. The key observation is that the self-model need not be completely accurate, or, rather, that there is a certain flexibility in what counts as an accurate answer, because what it says can't be contradicted by other sources of information. If everyone's self-model says they have free will, then free will can't be anything but whatever it is everyone thinks they have. It becomes difficult to deny that we have free will, because there's no content to the claim that we have it over and above what the chorus of self-models declare.[13]

Phenomenal experience now emerges as the self-model's answer to the question, What hap-

[13] For the complete story on free will, see (McDermott, 2001), ch. 3. I referred to Minsky's rather different theory above; McCarthy champions his own version in (McCarthy & Hayes, 1969).

pens when I perceive something? The answer, in terms of appearance, reality, and error, is accurate up to a point. It's when we get to qualia that the model ends the explanation with a just-so story. It gives more useful answers on questions such as whether it's easier to confuse green and yellow than green and red, or what to do when senses conflict, or what conditions make errors more or less likely. But to questions such as, How do I know this is red in the first place?, it gives an answer designed to stop inquiry. The answer is that red has this quality (please focus attention on the red object) which is intrinsically different from the analogous quality for green objects (now focus over here, if you don't mind). Because red is "intrinsically like … *this,*" there is no further question to ask. Nor should there be. I can take steps to improve my classification of objects by color, but there's nothing I can do to improve my ability to tell red from green (or, more plausibly, to tell two shades of red apart) once I've obtained optimal lighting and viewing conditions.[14]

The computationalist theory of phenomenal consciousness thus ends up looking like a spoilsport's explanation of a magic trick. It comes down to: "Don't look over there! The key move is over here, where you weren't looking!"[15] *Phenomenal consciousness is not part of the mechanism of perception, but part of the mechanism of introspection about perception.*

It is easy to think that this theory is similar to Perlis's model of self-consciousness as ultimately fundamental, and many philosophers have misread it that way. That's why "self-consciousness" is so misleading. Ordinarily what we mean by it is consciousness of self. But the self-model theory of consciousness aims to explain *all* phenomenal consciousness in terms of *subpersonal* modelling by an organism R of R's own perceptual system. Consciousness of self is just a particular sort of phenomenal consciousness, so the theory aims to explain it in terms of modelling by R of R's own perceptual system in the act of perceiving R. In these last two sentences the word "self" does not appear except as part of the *definiendum*, not as part of the *definiens*. Whatever the "self" is, it is not lying around waiting to be perceived; the act of modelling it defines what it is to a great extent. There is nothing mystical going on here. When R's only view of R is $R*$, in Minsky's terminology, then it is no surprise if terms occur in $R*$ whose meaning depends at least partly on how $R*$ fits into everything else R is doing, and in particular on how (the natural-language equivalents of those) terms are used by a community of organisms R belongs to.

I think the hardest part of this theory to accept is that perception is normally not mediated, or even accompanied, by qualia. In section 1 I invited readers to cast their eyes over a complex climate-control system and observe the absence of sensation. We can do the same exercise with the brain, with the same result. It just doesn't need sensations in order to do its job. But if you *ask* it, it will claim it does. A quale exists only when you look for it.

[14] One may view it as a bug that a concept, qualia, whose function is to end introspective questioning has stimulated so much conversation! Perhaps if human evolution goes on long enough natural selection will eliminate those of us who persist in talking about such things, especially while crossing busy streets.

[15] Cf (Wittgenstein, 1953): "The decisive movement in the conjuring trick has been made, and it was the very one we thought quite innocent."

Throughout this section, I have tried to stay close to what I think is a consensus position on a computational theory of phenomenal consciousness. But I have to admit that the endpoint to which I think we are driven is one that many otherwise fervent computationalists are reluctant to accept. There is no alternative conclusion on the horizon, just a wish for one, as in this quote from (Perlis, 1997):

> . . . Perhaps bare consciousness is in and of itself a self-distinguishing process, a process that takes note of itself. If so, it could still be considered a quale, the ur-quale, what it's like to be a bare subject. . . . What might this be? That is unclear. . . .

Perlis believes that a conscious system needs to be "strongly self-referring," in that its modelling of self is modelled in the very modelling, or something like that. "Why do we need a self-contained self, where referring stops? Negotiating one's way in a complex world is tough business. . . ." He sketches a scenario in which Ralph, a robot, needs a new arm.

> Suppose the new arm is needed within 24 hours. He cannot allow his decision-making about the best and quickest way to order the arm get in his way, i.e., he must not allow it to run on and on. He can use meta-reasoning to watch his reasoning so it does not use too much time, but then what is to watch his meta-reasoning? . . . He must budget his time. Yet the budgeting is another time-drain, so he must pay attention to that too, and so on in an infinite regress. . . . Somehow he must regard [all these modules] as himself, one (complex) system reasoning about itself, *including that very observation.* He must *strongly self-refer*: he must refer to that very referring so that its own time-passage can be taken into account. (Italics in original.)

It appears to me that two contrary intuitions are colliding here. One is the hard-headed computationalist belief that self-modelling is all you need for consciousness; the other is the nagging feeling that self-modelling alone can't quite get us all the way. Yet when he tries to find an example, he winds up with a mystical version of the work by Boddy and Dean (1989) that I cited above as a *prosaic* example of self modelling. It seems clear to me that the only reason Perlis needs the *virtus dormitiva* of "strong self-reference" is because the problem-solving system he's imagining is not an ordinary computer program, but a transcendental self-contemplating mind, something not really divided into modules at all, but actively dividing itself into time-shared virtual modules as it shifts its attention from one aspect of its problem to another, then to a meta-layer, a meta-meta-layer, and so forth. If you bite the bullet and accept that all this meta-stuff, if it exists at all, exists only in the system's self-model, then the need for strong self-reference, and the "ur-quale," goes away, much like the ether in the theory of electromagnetism. So I believe, but I admit that most AI researchers who take a position probably share Perlis's reluctance to let that ether go.

5 The Critics

AI has always generated a lot of controversy. The typical pattern is that some piece of research captures the public's imagination, as amplified by journalists, then the actual results don't fit those public expectations, and finally someone comes along to chalk up one more failure of AI research. Meanwhile, often enough the research does succeed, not on the goals hallucinated by the popular press, but on those the researchers actually had in mind, so that the AI community continues to gain confidence that it is on the right track.

Criticism of AI models of consciousness doesn't fit this pattern. As I observed at the outset, almost no one in the field is "working on" consciousness, and certainly there's no one trying to write a conscious program. It is seldom that a journalist can make a breathless report about a *robot that will actually have experiences!!*[16]

Nonetheless, there has been an outpouring of papers and books arguing that mechanical consciousness is impossible, and that suggestions to the contrary are wasteful of research dollars and possibly even dangerously dehumanizing. The field of "artificial consciousness" (AC) is practically *defined* by writers who deny that such a thing is possible. Much more has been written by AC skeptics than by those who think it is possible. In this section I will discuss some of those criticisms and refute them as best I can.

Due to space limitations, I will try to focus on critiques that are specifically directed at computational models of consciousness, as opposed to general critiques of materialist explanation. For example, I will pass over Jackson's (1982) story about "Mary, the color scientist" who learns what red looks like. There are interesting things to say about it (which I say in (McDermott, 2001)), but Jackson's critique is not directed at, and doesn't mention, computationalism in particular. I will also pass over the vast literature on "inverted spectrum" problems, which is a somewhat more complex version of the sour/spicy taco problem.

Another class of critiques that I will omit are those whose aim is to show that computers can never achieve human-level intelligence. As discussed in sections 3, I concede that if computers can't be intelligent then they can't be conscious either. But our focus here is on consciousness, so the critics I try to counter are those who specifically argue that computers will never be conscious, even if they might exhibit intelligent behavior. One important group of arguments this leaves out are those based on Gödel's proof that Peano arithmetic is incomplete (Nagel & Newman, 1958; Penrose, 1989, 1994). These arguments are intended to show a limitation in the abilities of computers to reason, not specifically a limitation on their ability to experience things; in fact, the connection between the two is too tenuous to justify talking about the topic in detail.

[16] One occasionally hears news reports about attempts to build an artificial nose. When I hear such a report, I picture a device that measures concentrations of substances in the air. But perhaps the average person imagines a device that "smells things," so that, e.g., the smell of a rotten egg would be unpleasant for it. In any case, these news reports seem not to have engendered much controversy, so far.

5.1 Turing's Test

Let's start where the field started: with Turing's Test (Turing, 1950). As described in section 3, the test consists of a judge trying to distinguish a computer from a person by carrying on typed conversations with them. If the judge gets it wrong about 50% of the time, then the computer passes the test.

Turing's test is not necessarily relevant to the computational theory of consciousness. Few of the theorists discussed in sections 3 and 4 have invoked the Test as a methodological tool. Where it comes in is when it is *attributed* to computationalists. A critic will take the computationalist's focus on the third-person point of view as an endorsement of behaviorism, then jump to Turing's Test as the canonical behaviorist tool for deciding whether an entity is conscious. That first step, from "third-person" to "behaviorist," is illegitimate. It is, in fact, somewhat ludicrous to accuse someone of being a behaviorist who is so eager to open an animal up (metaphorically, that is), and stuff its head with intricate block diagrams. All the "third-personist" is trying to do is stick to scientifically, that is, publicly, available facts. This attempt is biased against the first-person view, and that bias pays off by eventually giving us an *explanation* of the first person.

So there is no particular reason for a computationalist to defend the Turing Test. It doesn't particularly help develop theoretical proposals, and it gets in the way of thinking about intelligent systems that obviously can't pass the test. Nonetheless, an objection to computationalism raised in section 3.1 does require an answer. That was the objection that even if a computer could pass the Turing Test, it wouldn't provide any evidence that it actually was conscious. I disagree with this objection on grounds that should be clear at this point: To be conscious is to model one's mental life in terms of things like sensations and free decisions. It would be hard to have an intelligent robot that wasn't conscious in this sense, because everywhere the robot went it would have to deal with its own presence and its own decision making, and so it would have to have models of its behavior and its thought processes. Conversing with it would be a good way of finding out how it thought about itself, that is, what its self-models were like.

Keep in mind, however, that the Turing Test is not likely to be the standard method to check for the presence of consciousness in a computer system, if we ever need a standard method. A robot's self-model, and hence its consciousness, could be quite different from ours in respects that are impossible to predict given how far we are from having intelligent robots. It is also just barely possible that a computer not connected to a robot could be intelligent with only a very simple self-model. Suppose the computer's job was to control the traffic, waste management, and electric grid of a city. It might be quite intelligent, but hardly conscious in a way we could recognize, simply because it wouldn't be present in the situations it modeled the way we are. It probably couldn't pass the Turing Test either.

Somewhere in this thicket of possibilities there might be an artificial intelligence with an alien form of consciousness that could *pretend* to be conscious on our terms while knowing full well that it wasn't. It could then pass the Turing Test, wine tasting division, by faking it. All this shows is that there is a slight possibility that the Turing Test could be good at detecting intelligence and not so good at detecting consciousness. This shouldn't give much comfort to those who think that the Turing Test systematically distracts us from the first-person viewpoint. If someone ever builds a machine that passes it, it will certainly exhibit intentionality and intelligence, and almost certainly be conscious. There's a remote chance that human-style consciousness can be faked, but no chance that intelligence can be.[17]

5.2 The Chinese Room

One of the most notorious arguments in the debate about computational consciousness is Searle's (1980) "Chinese Room" argument. It's very simple. Suppose we hire Searle (who speaks no Chinese) to implement a computer program for reading stories in Chinese and then answering questions about those stories. Searle reads each line of the program and does what it says. He executes the program about a million times slower than an actual CPU would, but if we don't mind the slow motion we could carry on a perfectly coherent conversation with him.

Searle goes on:

> Now the claims made by strong AI are that the programmed computer understands the stories and that the program in some sense explains human understanding. But we are now in a position to examine these claims in light of our thought experiment.
>
> 1. As regards the first claim, it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. . . .
>
> 2. As regards the second claim, that the program explains human understanding, we can see that the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding.

It's hard to see what this argument has to do with consciousness. The connection is somewhat indirect. Recall that in section 4.2 we made sure to talk about "impersonal" intentionality, the kind a system has by virtue of being a computer whose symbol structures are causally connected to the environment so as to denote objects and states of affairs in that environment. Searle absolutely refuses to grant that there is any such thing as impersonal or subpersonal intentionality (Searle, 1992). The paradigm case of any mental state is always the conscious

[17] I realize that many people, for instance Robert Kirk (1994), believe that in principle something as simple as a lookup table could simulate intelligence. I don't have space here to refute this point of view, except to note that besides the fact that the table would be larger than the known universe and take a trillion years to build, a computer carrying on a conversation by consulting it would not be able to answer a question about what time it is.

mental state, and he is willing to stretch mental concepts only far enough to cover unconscious mental states that could have been conscious (repressed desires, for instance). Hence there is no understanding of Chinese unless it is accompanied by a conscious awareness or feeling of understanding.

If Searle's stricture were agreed upon, then all research in cognitive science would cease immediately, because it routinely assumes the existence of nonconscious symbol processing to explain the results of experiments.[18]

Searle seems to have left an escape clause, the notion of "weak AI":

> I find it useful to distinguish what I will call 'strong' AI from 'weak' or 'cautious' AI .... According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states (Searle, 1980).

Many people have adopted this terminology, viewing the supposed weak version of AI as a safe harbor in which to hide from criticism. In my opinion, the concept of weak AI is incoherent. Suppose someone writes a program to simulate a hurricane, to use a common image. The numbers in the simulation denote actual or hypothetical air pressures, wind velocities, and the like. The simulation embodies differential equations that are held to be more or less true statements about how wind velocities affect air pressures and vice versa, and similarly for all the other variables involved. Now think about "computer simulations of human cognitive capacities" (Searle's phrase). What are the analogues of the wind velocities and air pressures in this case? When we use the simulations to "formulate and test hypotheses," what are the hypotheses *about*? They might be about membrane voltages and currents in neurons, but of course they aren't, because neurons are "too small." We would have to simulate an awful lot of them, and we don't really know how they're connected, and the simulation would just give us a huge chunk of predicted membrane currents anyway. So no one does that. Instead, they run simulations at a much higher level, at which symbols and data structures emerge. This is true even for neural-net researchers, whose models are much, much smaller than the real thing, so that each connection weight represents an abstract summary of a huge collection of real weights. What, then, is the ontological status of these symbols and data structures? If we believe that these symbols and the computational processes over them are really present in the

[18] There is a popular belief that there is such a thing as "nonsymbolic" or "subsymbolic" cognitive science, as practiced by those who study artificial neural nets. As I mentioned in section 4.1, this distinction is usually unimportant, and the present context is an example. The goal of neural-net researchers is to explain conscious thought in terms of unconscious computational events in neurons, and as far as Searle is concerned, this is just the same fallacy all over again (Searle, 1990).

brain, and really explain what the brain does, then we are back to strong AI. But if we don't believe that, then why the hell are we simulating them? By analogy, let us compare strong vs. weak computational meteorology. The former is based on the belief that wind velocities and air pressures really have something to do with how hurricanes behave. The latter allows us to build "powerful tools" that perform "computer simulations of [hurricanes' physical] capacities," and "formulate and test hypotheses" about . . . something *other* than wind velocities and air pressures?

Please note that I am not saying that all cognitive scientists are committed to a computationalist account of consciousness. I'm just saying that they're committed to a computationalist account of whatever it is they're studying. If someone believes that the EPAM model (Feigenbaum & Simon, 1984) accounts for human errors in memorizing lists of nonsense syllables, they have to believe that structures isomorphic to the discrimination trees in EPAM are actually to be found in human brains. If someone believes that there is *no* computationalist account of consciousness, then they must also believe that a useful computer simulation of consciousness must simulate something *other* than symbol manipulation, perhaps ectoplasm secretions. In other words, given our lack of *any* noncomputational account of the workings of the mind, they must believe it to be pointless to engage in simulating consciousness *at all* at this stage of the development of the subject.

There remains one opportunity for confusion. No one believes that a simulation of a hurricane could blow your house off the beach. Why should we expect a simulation of a conscious mind to be conscious (or expect a simulation of a mind to be a mind)? Well, we need not expect that, exactly. If a simulation of a mind is disconnected from an environment, then it would remain a mere simulation.

However, once the connection is made properly, we confront the fact that a sufficiently detailed simulation of computation C *is* computation C. This is a property of formal systems generally. As Haugeland (1985) observes, the difference between a game like tennis and a game like chess is that the former involves moving a physical object, the ball, through space, while the latter involves jumping from one legal board position to the next, and legal board positions are not physical entities. In tennis, one must hit a ball with certain prescribed physical properties using a tennis racket, which must also satisfy certain physical requirements. Chess requires only that the state of the game be represented with enough detail to capture the positions of all the pieces.[19] One can use any 8×8 array as a board, and any collection of objects as pieces, provided they are isomorphic to the standard board and pieces. One can even use computer data structures. So a detailed simulation of a good chess player *is* a good chess player, provided it is connected by some channel, encoded however you like, between its computations and an actual opponent with whom it is alternating moves. Whereas for a simulation of a tennis player to be a tennis player, it would have to be connected to a robot capable of tracking and hitting tennis balls.

[19] And a couple of other bits of information, such as whether each player still has castling as an option.

This property carries over to the simulation of any other process that is essentially computational. So, if it happens that consciousness is a computational phenomenon, then a sufficiently faithful simulation of a conscious system would be a conscious system, provided it was connected to the environment in the appropriate way. This point is especially clear if the computations in question are somewhat modularizable, as might be the case for a system's self-model. The difference between a nonconscious tennis player and a conscious one might involve connections among its internal computational modules, and not the connections from there to its cameras and motors. There would then be no difference between the "consciousness module" and a detailed simulation of that "module"; they would be interchangeable, provided that they didn't differ too much in speed, size, and energy consumption. I use scare quotes here because I doubt that things will turn out to be that tidy. Nonetheless, no matter how the wires work out, the point is that nothing *other than* computation need be involved in consciousness, which is what Strong AI boils down to. Weak AI boils down to a sort of "cargo cult" whose rituals involve simulations of things someone only guesses might be important in some way.

Now that I've clarified the stakes, let's look at Searle's argument. It is ridiculously easy to refute. When he says, "the claims made by strong AI are that the programmed computer understands the stories and that the program in some sense explains human understanding," he may be right about the second claim (depending on how literally you interpret "explains"), but he is completely wrong about the first claim, that the programmed computer understands something. As McCarthy says, "The Chinese Room Argument can be refuted in one sentence: Searle confuses the mental qualities of one computational process, himself for example, with those of another process that the first process might be interpreting, a process that understands Chinese, for example" (McCarthy, 2000). Searle's slightly awkward phrase "the programmed computer" gives the game away. Computers and software continually break our historical understanding of the identity of objects across time. Any computer user has (too often) had the experience of not knowing "whom" they're talking to when talking to their program. Listen to a layperson try to sort out the contributions to their current state of frustration of the e-mail delivery program, the e-mail reading program, and the e-mail server. When you run a program you usually then talk to it. If you run two programs at once you switch back and forth between talking to one and talking to the other.[20] The phrase "programmed computer" makes it sound as if programming it changes *it* into something you can talk to. The only reason to use such an odd phrase is because in the story Searle himself plays the role of the programmed computer, the entity that doesn't understand. By pointing at the "human CPU" and shouting loudly, he hopes to distract us from the abstract entity that is brought into existence by executing the story-understanding program.

We can state McCarthy's argument vividly by supposing that *two* CPUs are involved, as they might well be. The story-understanding program might be run on one for a while, then

[20] Technically I mean "process" here, not "program." McCarthy's terminology is more accurate. But I'm trying to be intelligible by technical innocents.

on the other, and so forth, as dictated by the internal economics of the operating system. Do we imagine that the ability to "understand" jumps back and forth between the two CPUs? If we replace the two CPUs by two people, does Strong AI predict that the ability to understand Chinese will jump back and forth between the two people (McDermott, 2001)? Of course not.

5.3 Symbol Grounding

In both of the preceding sections, it sometimes seems as if intentionality is the real issue, or what Harnad (1990, 2001) calls the *symbol-grounding problem.* The problem arises from the idea of a disembodied computer living in a realm of pure syntax, which we discussed in section 3.5. Suppose that such a computer ran a simulation of the battle of Waterloo. That is, we intend it to simulate that battle, but for all we know there might be another encoding of its states that would make it be a simulation of coffee prices in Ecuador.[21] What connects the symbols to the things they denote? In other words, what *grounds* the symbols?

This problem underlies some people's concerns about the Turing Test and the Chinese Room because the words in the Turing Test conversation might be considereed to be ungrounded and therefore meaningless (Davidson, 1990); and the program and data structures being manipulated by the human CPU John Searle seem also to be disconnected from anything that could give them meaning.

As should be clear from the discussion in section 4.2, symbols get their meanings by being causally connected to the world. Harnad doesn't disagree with this, but he thinks that the connection must take a special form, via neural networks, natural or artificial.[22] The inputs to the networks must be sensory transducers. The outputs are neurons that settle into different stable patterns of activation depending on how the transducers are stimulated. The possible stable patterns, and the way they classify inputs, is learned over time as the network is trained by its owner's encounters with it surroundings.

> How does the hybrid system find the invariant features of the sensory projection that make it possible to categorize and identify objects correctly? Connectionism, with its general pattern-learning capability, seems to be one natural candidate (though there may well be others): Icons, paired with feedback indicating their names, could be processed by a connectionist network that learns to identify icons correctly from the sample of confusable alternatives it has encountered by dynamically adjusting the weights of the features and feature combinations that are reliably associated with the names in a way that (provisionally) resolves the confusion, thereby reducing the icons to the invariant (confusion-resolving) features of the category to which they are assigned. In effect, the 'connection' between the names

[21] I believe these particular examples (Waterloo & Ecuador) were invented by someone besides me, but I have been unable to find the reference.

[22] The fact that these are called "connectionist" is a mere pun in this context — I hope.

> and the objects that give rise to their sensory projections and their icons would be provided by connectionist networks (Harnad, 1990).

The symbol-grounding problem, if it is a problem, requires no urgent solution, as far as I can see. I think it stems from a basic misunderstanding about what computationalism is and what the alternatives are. Harnad's view is "The predominant approach to cognitive modeling is still what has come to be called 'computationalism' . . . , the hypothesis that cognition is computation. The more recent rival approach is 'connectionism' . . . , the hypothesis that cognition is a dynamic pattern of connections and activations in a 'neural net' " (Harnad, 2001). Put this way, it seems clear that neural nets would be allowed under computationalism's "big tent," but Harnad withdraws the invitation rapidly, by imposing a series of fresh requirements. By "computation" he means "symbolic computation," which consists of syntactic operations on "symbol tokens." Analogue computation is ruled out. Symbolic computation doesn't depend on the medium in which it is implemented, just so long as it is implemented somehow (because the syntactic categories of the symbol tokens will be unchanged). And last, but certainly not least, "the symbols and symbol manipulations in a symbol system [must be] systematically interpretable (Fodor & Pylyshyn, 1988): they can be assigned a semantics, they mean something (e.g., numbers, words, sentences, chess moves, planetary motions, etc.)." The alternative is "trivial" computation, which produces "uninterpretable formal gibberish."

As I argue in (McDermott, 2001), these requirements have seldom been met by what most people call "computational" systems. The average computer programmer knows nothing about formal semantics or systematic interpretability. Indeed, in my experience it is quite difficult to teach a programmer about formal systems and semantics. One must scrape away layers of prior conditioning about how to "talk" to computers.

Furthermore, as I said in section 4.1, few AI practitioners refuse to mix and match connectionist and symbolic programs. One must be careful about how one interprets what they *say* about their practice. Clancey (1999), in arguing for a connectionist architecture, calls the previous tradition modeling the brain as a "wet" computer similar in important respects to the "dry" computers we use as models. He argues that we should replace it with a particular connectionist architecture. As an example of the change this would bring, he says (p. 30) "Cognitive models have traditionally treated procedural memory, including inference rules ('if X then Y'), as if human memory is just computer random-access memory. . . ." He proposes to "explore the hypothesis that a sequential association, such as an inference rule . . . , is a temporal relation of activation, such that if X implies Y," what is recorded is a "relation . . . of temporal activation, such that when X is presently active, Y is a categorization that is potentially active next" (p. 31). But he remains a committed computationalist through this seemingly discontinuous change. For instance, in discussing how the new paradigm would actually work, he says "The discussion of [insert detailed proposal here] illustrates how the discipline of implementing a process in a computer representation forces distinctions to be rediscovered and brings into question consistency of the theory" (p. 44).

The moral is that we must be careful to distinguish between two ways computers are used in psychological modelling: as implementation platform and as metaphor. The digital-computer metaphor might shed light on why we have a single stream of consciousness ($\sim$ von Neumann instruction stream?), why we can only remember 7 ± 2 things ($\sim$ size of our register set?), why we have trouble with deep center-embedded sentences like "The boy the man the dog bit spanked laughed" ($\sim$ stack overflow?). The metaphor may have had some potential in the 1950s, when cognitive science was just getting underway, but it's pretty much run out of steam at this point. Clancey is correct to point out how the metaphor may have affected cognitive science in ways that seemed too harmless to notice, but that in retrospect are hard to justify. For instance, the program counter in a computer makes pursuing a rigid list of tasks easy. If we help ourselves to a program counter in implementing a cognitive model, we may have begged an important question about how sequentiality is achieved in a parallel system like the brain.

What I argue is that the essence of computationalism is to believe (a) that brains are essentially computers; and (b) digital computers can simulate them in all important respects, even if they aren't digital at all. Because a simulation of a computation *is* a computation, the "digitality" of the digital computer cancels out. If symbol grounding is explained by some very special properties of a massively parallel neural network of a *particular sort*, then if that net can be simulated in real time on a cluster of parallel workstations, then the cluster becomes a virtual neural net, which grounds symbols as well as a "real" one would.

Perhaps this is the place to mention the paper by O'Brien & Opie (1999) that presents a "connectionist theory of phenomenal experience." The theory makes a basic assumption, that a digital simulation of a conscious connectionist system would not be conscious. It is very hard to see how this could be true. It's the zombie hypothesis, raised from the dead one more time. The "real" neural net is conscious, but the simulated one, in spite of behaving in exactly the same way (plus or minus a little noise), would be experience-less — another zombie lives.

6 Conclusions

The contribution of artificial intelligence to consciousness studies has been slender so far, because almost everyone in the field would rather work on better defined, less controversial problems. Nonetheless, there do seem to be common themes running through the work of AI researchers that touches on phenomenal consciousness. Consciousness stems from the structure of the *self-models* that intelligent systems use to reason about themselves. A creature's models of itself are like models of other systems, except for some characteristic indeterminacy about what counts as accuracy. In order to explain how an information-processing system can *have* a model of something, there must be a prior notion of intentionality that explains why and how symbols inside the system can refer to things. This theory of *impersonal intentionality* is based on the existence of harmonious matchups between the states of the system and states of the world. The meanings of symbol structures are what the matchups say they are.

Having established that a system's model of that very system is a nonvacuous idea, the next step is to show that the model almost certainly will contain ways of thinking about how the system's senses work. The difference between appearance and reality arises at this point, and allows the system to reason about its errors in order to reduce the chance of making them. But the self-model also serves to set boundaries to the questions that it can answer. The idea of a sensory quale arises as a useful way of cutting off useless introspection about how things are ultimately perceived and categorized.

Beyond this point it is hard to find consensus between those who believe that the just-so story the self-model tells its owner is all you need to explain phenomenal consciousness, and those who think that something more is needed. Frustratingly, we won't be able to create systems and test hypotheses against them in the foreseeable, because real progress on creating conscious programs awaits further progress on enhancing the intelligence of robots. There is no guarantee that AI will *ever* achieve the requisite level of intelligence, in which case this chapter has been pretty much wasted effort.

There are plenty of critics who don't want to wait to see how well AI succeeds, because they think they have arguments that can shoot down the concept of machine consciousness, or rule out certain forms of it, right now. We examined three: the accusation that AI is behaviorist on the subject of consciousness, the "Chinese Room" argument, and the symbol-grounding problem. In each case the basic computationalist working hypothesis survived intact: that the embodied brain is an "embedded" computer, and that a reasonably accurate simulation of it would have whatever mental properties it has, including phenomenal consciousness.

References

Baars, B. J. (1988). *A cognitive theory of consciousness.* New York: Guilford Press.

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind.* New York: Oxford University Press.

Block, N. (1997). On a confusion about a function of consciousness. In Block et al. (1997), pp. 375–415.

Block, N., Flanagan, O., & Güzeldere, G. (eds) (1997). *The nature of consciousness: Philosophical debates.* Cambridge, Mass.: MIT Press.

Boddy, M. & Dean, T. (1989). Solving time-dependent planning problems. *Proc. Ijcai*, *11*, pp. 979–984.

Campbell, J. (1994). *Past, space and self.* MIT Press: Cambridge.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory.* New York: Oxford University Press.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain.* Cambridge, Mass.: MIT Press.

Churchland, P. M. (1988). *Matter and consciousness: A contemporary introduction to the philosophy of mind.* Cambridge, Mass.: MIT Press.

Clancey, W. J. (1999). *Conceptual coordination: How the mind orders experience in time.* New Jersey: Lawrence Erlbaum Associates. Mahwah.

Currie, K. & Tate, A. (1991). O-plan: the open planning architecture. *Artificial Intelligence*, *52*(1), pp. 49–86.

Davidson, D. (1990). Turing's test. In Said et al. (1990), pp. 1–11.

Dawkins, R. (1976). *The selfish gene.* Oxford: Oxford University Press.

Dennett, D. C. (1969). *Content and consciousness.* London: Routledge & Kegan Paul. International Library of Philosophy and Scientific Method.

Dennett, D. C. (1991a). Real patterns. *Journal of Philosophy*, *88*, pp. 27–51.

Dennett, D. C. (1991b). *Consciousness explained.* Boston: Little, Brown and Company.

Douglas, G. & Saunders, S. (2003). Dan Dennett: Philosopher of the month. *TPM Online: the Philosophers' Magazine on the internet.* Retrieved December 31, 2005 from
`http://www.philosophers.co.uk/cafe/phil_apr2003.htm`.

Dretske, F. I. (1981). *Knowledge and the flow of information.* Cambridge, Mass.: MIT Press.

Feigenbaum, E. A. & Simon, H. A. (1984). Epam-like models of recognition and learning. *Cognitive Sci*, *8*(4), pp. 305–336.

Fodor, J. (1975). *The language of thought.* New York: Thomas Y. Crowell.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. In Pinker & Mehler (1988), pp. 3–72.

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

Harnad, S. (2001). Grounding symbols in the analog world with neural nets – a hybrid model. *Psycoloquy.* Retrieved December 31, 2005 from
`http://psycprints.ecs.soton.ac.uk/archive/00000163/`.

Haugeland, J. (1985). *Artificial intelligence: The very idea.* Cambridge, Mass.: MIT Press.

Hayes-Roth, B. (1985). A blackboard architecture for control. *Artificial Intelligence*, *26*(3), pp. 251–321.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid.* New York: Basic Books.

Hofstadter, D. R. & Dennett, D. C. (1981). *The mind's I: Fantasies and reflections on self and soul.* New York: Basic Books.

Jackson, F. (1982). Epiphenomenal qualia. *Phil. Quart.*, *32*, pp. 127–136.

Jaynes, J. (1976). *The origins of consciousness in the breakdown of the bicameral mind.* Boston: Houghton Mifflin.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, MA: Harvard University Press.

Kirk, R. (1994). *Raw feeling: A philosophical account of the essence of consciousness.* Oxford: Oxford University Press.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence.* New York: Penguin Books.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*, pp. 529–566.

McCarthy, J. (1990a). *Formalizing common sense.* Norwood, NJ: Ablex.

McCarthy, J. (1990b). Ascribing mental qualities to machines. In McCarthy (1990a).

McCarthy, J. (1995a). Todd Moody's zombies. *J. Consciousness Studies.* Retrieved December 31, 2005 from `http://www-formal.stanford.edu/jmc/zombie/zombie.html`.

McCarthy, J. (1995b). Making robots conscious of their mental states. *Proc. Machine Intelligence Workshop.* Retrieved December 31, 2005 from `http://www-formal.stanford.edu/jmc/consciousness.html`.

McCarthy, J. (2000). John Searle's Chinese room argument. Retrieved December 31, 2005 from `http://www-formal.stanford.edu/jmc/chinese.html`.

McCarthy, J. & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer & Michie (1969), pp. 463–502.

McDermott, D. (2001). *Mind and mechanism.* Cambridge, Mass.: MIT Press.

Meltzer, B. & Michie, D. (eds) (1969). *Machine intelligence 4.* Edinburgh University Press.

Minsky, M. (1968a). *Semantic information processing.* Cambridge, Mass: MIT Press.

Minsky, M. (1968b). Matter, mind, and models. In Minsky (1968a), pp. 425–432.

Minsky, M. (1986). *The society of mind.* New York: Simon and Schuster.

Moody, T. C. (1994). Conversations with zombies. *J. Consciousness Studies*, *1*(2), pp. 196–200.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, *38*(8), pp. 114–117.

Moravec, H. P. (1988). Sensor fusion in certainty grids for mobile robots. *AI Magazine*, *9*, pp. 61–74. Summer 88.

Moravec, H. (1999). *Robot: Mere machine to transcendent mind.* New York: Oxford University Press.

Nagel, E. & Newman, J. R. (1958). *Goedel's proof.* New York: New York University Press.

O'Brien, G. & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*, pp. 127–148.

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics.* New York: Oxford University Press.

Penrose, R. (1994). *Shadows of the mind: A search for the missing science of consciousness.* New York: Oxford University Press.

Perlis, D. (1994). An error-theory of consciousness. U. of Maryland Computer Science. CS-TR-3324.

Perlis, D. (1997). Consciousness as self-function. *J. Consciousness Studies*, *4*(5/6), pp. 509–25.

Pinker, S. & Mehler, J. (1988). *Connections and symbols.* Cambridge, Mass.: MIT Press.

Rey, G. (1997). A question about consciousness. In Block et al. (1997), pp. 461–482.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, Mass.: MIT Press.

Said, K. M., Newton-Smith, W., Viale, R., & Wilkes, K. (1990). *Modelling the mind.* Oxford: Clarendon Press.

Scheutz, M. (ed) (2002). *Computationalism: New directions.* Cambridge, Mass.: The MIT Press.

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

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

Searle, J. R. (1992). *The rediscovery of the mind.* Cambridge, Mass.: MIT Press.

Sloman, A. (1978). *The computer revolution in philosophy.* Hassocks, Sussex: The Harvester Press.

Sloman, A. (2002). The irrelevance of Turing machines to artificial intelligence. In Scheutz (2002), pp. 87–127.

Sloman, A. & Chrisley, R. (2003). Virtual machines and consciousness. *J. Consciousness Studies*, *10*(4–5), pp. 6–45.

Smith, B. C. (1984). Reflection and semantics in Lisp. *Proc. Conf. on Principles of Programming Languages*, *11*, pp. 23–35.

Smith, B. C. (1995). *On the origins of objects.* Cambridge, Mass.: MIT Press.

Smith, B. C. (2002). The foundations of computing. In Scheutz (2002), pp. 23–58.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *49*, pp. 433–460.

Wilkes, K. (1990). Modelling the mind. In Said et al. (1990), pp. 63–82.

Winograd, T. (1972). *Understanding natural language.* New York: Academic Press.

Wittgenstein, L. (1953). *Philosophical investigations.* New York: The MacMillan Company.