Real Robots and the Missing Thought Experiment in the Chinese Room Dialectic*

Selmer Bringsjord & Ron Noel
Dept. of Philosophy, Psychology & Cognitive Science
Department of Computer Science (S.B.)
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180-3590 USA
selmer@rpi.edu • noelr@rpi.edu

December 26, 2000

Introduction

John Searle's [19], [20] Chinese Room Argument (CRA) is arguably the 20^{th} century's greatest philosophical polarizer. On the one hand, as the years pass, his argument is looking more and more to be headed for philosophical immortality. (Consider the book you're holding, published a full two decades after the argument's debut.) On the other hand, a common attitude among Strong AIniks is that CRA is not only unsound, but silly, based as it is on a fanciful story (CR) far removed from the *practice* of AI — practice which is year by year moving ineluctably toward sophisticated robots that will once and for all silence CRA and its proponents. For example, John Pollock, a philosopher and Strong AInik who takes his OSCAR system to be a significant step toward such robots, writes:

> Once OSCAR is fully functional, the argument from analogy will lead us inexorably to attribute thoughts and feelings to OSCAR with precisely the same credentials with which we attribute them to human beings. Philosophical arguments to the contrary will be passé. ([18], 6)

The position that smarter and smarter robots will kill off Searle's CRA (and other arguments against Strong AI) is also expressed in Hans Moravec's recent book, *Robot: Mere Machine to Transcendent Mind* [17]. Moravec cites Searle's CRA, and then simply goes on to describe a (near) future in which robots do everything we do, and more — an age in which it is taken for granted that robots have what CRA says they cannot have: subjective conscious states. There are many others who feel that the advance of robots will demote CRA to the lowly ranks of arguments like those given in the past for such propositions as that flight is impossible and that the Earth is flat. For example, in personal conversation with one of us (Bringsjord), Pat Hayes, John McCarthy, and Marvin Minsky — all three of whom have been involved in real robot building since AI's inception — have said outright that CRA is silly, and that as robots get smarter and smarter, this argument will wither

*For helpful comments on and criticisms of elements of this chapter and its ancestors, we are indebted to David Chalmers, Jack Copeland, Eric Dietrich, Jim Fetzer, Stevan Harnad, Pay Hayes, Marvin Minsky, Jim Moor, and John Searle. For the engineering work that was required to make the "zombanimal" robots described herein a reality, we're indebted to Clarke Caporale.

away to a quaint curiosity. In particular, Hayes has said that Searle and like-minded thinkers are specifically analogous to those learned but ludicrous-in-hindsight men who declared that flight was impossible.

In this chapter we attempt some philosophical jujitsu against those who think real robots spell trouble for CRA: we show how real robots can be used to *strengthen* CRA.

We begin by refuting the modernized version of the Robot Reply to CRA, which is due to Stevan Harnad and Daniel Dennett; this modernized reply stands at present unscathed, despite Searle's own rejoinder to it. Part of our refutation is enabled by a new thought-experiment, one we call (for reasons to be explained) "the missing thought-experiment" — in which figure robots and robotic appendages built and configured in our Minds & Machines Laboratory. This thought-experiment seems to describe not just a logical possibility, but a *physical* possibility. If we are right about this, it would seem that Searle has laid the foundation for demonstrating that the progress AI is making in building robots is merely progress in building "zombie animals," or, as we call them for short, "zombanimals."[1] If this is right, then CRA will forever live despite the prowess of robots. To put it as one of has before [2], robots will *do* a lot, maybe even as much as we do — play invincible chess, debate the finer points of philosophy of mind, drive race cars, fly jets, teach by the Socratic method, ski, write sustained philosophical arguments, and so on. But despite all this, robots won't *be* a lot: they won't be persons; they'll just *look* like persons. It may even be fair to say — once again, *if* we're right — that AI and robotics, viewed as enterprises aimed at building *minds* or *persons*, are futile to the point of being, we dare say, silly.

Our plan, specifically, is as follows. In section 1 we present the scheme presupposed by our coming analysis and argumentation. In section 2, using the scheme of section 1, we represent the original Robot Reply, the original System Reply, and Searle's responses to them. In section 3 we present and refute the modernized robot reply to CRA. In section 4 we first explain that the modernized robot reply, though fallacious, flirts with a move that *does* seem to threaten CRA — the move of *at once* appealing to the original System Reply *and* the original Robot Reply. We next point out that the upshot of this move, as noted by Dennett all the way back in his original commentary on CRA [11], is to create a demand for "the missing thought-experiment:" a scenario born of modifying CR so that the resultant version of CRA is at once immune from *both* the original Robot Reply *and* the original System Reply. Next, we briefly discuss Searle's truncated search for the missing thought-experiment. This discussion is followed by one devoted to the aforementioned zombanimals created in our Minds & Machines Laboratory. These silicon-based creatures are at the heart of the final stage of section 4, in which we provide the missing thought-experiment. As will soon be evident, we rely heavily on diagrams to help convey the various thought-experiments that form the heart of this chapter. In particular, our final diagram encapsulates the "missing thought-experiment" to which the title of this chapter refers.

1 The Scheme

There is no need to recapitulate, let alone formalize, CRA; it suffices here to establish merely a tidy "meta-perspective" on the argument and related issues.[2] To begin, we assume that CRA is a

[1] When we refer to zombies, we have in mind *philosophers'* zombies, not those creatures who shuffle about half-dead in the movies. (Actually, the zombies of cinematic fame apparently have real-life correlates created with a mixture of drugs and pre-death burial: see [9], [10].) We confine our attention in this chapter to the logical and physical possibility of zombanimals, rather than full-fledged zombies, mindless creatures who are indistinguishable from us. Bringsjord has elsewhere discussed such creatures at length [4].

[2] It's a trivial matter to express a formally valid version of CRA, as even Copeland (in his contribution to this volume) concedes, after presenting an invalid version of CRA. For a logically valid version of a CRA-like argument,

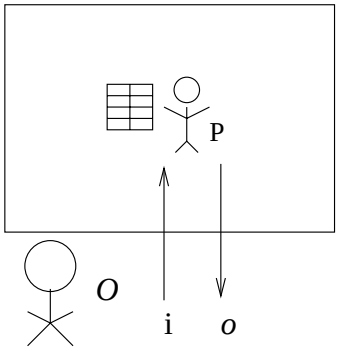

Figure 1: Pictorial Schema for Original Chinese Room, i.e., CR_1.

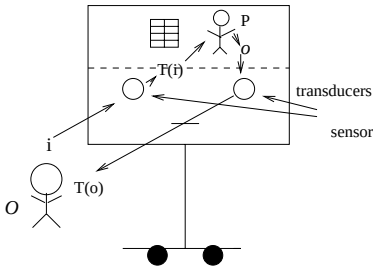

Figure 2: Pictorial Schema of Searle's Response (CR_2) to the Original Robot Reply (RR_1). Searle is inside the robot's cranium and the robot's transducers provide the interchange between input i from the native Chinese speakers (O), and the output o sent from Searle, which is transduced to $T(o)$ and then given to O.

deductive argument whose conclusion is the denial of the proposition that[3]

see the chapter "Searle" from Bringsjord's *What Robots Can and Can't Be* [2]. The analogue to what Copeland calls the 'Part-Of' principle in Bringsjord's version is the claim that if a human person P simulates n language programs (for Chinese, Norwegian, Spanish, etc.) it is wrong (indeed, silly!) to say that there are n different persons (n different bearers of the phenomenal consciousness that accompanies understanding conversation in natural language) that pop into existence.

[3]The modal necessity operator in (CC) — and in (TT), a proposition given just below — may strike many readers as too strong. One of us (Bringsjord) has discussed this and related issues in detail elsewhere [3], [6], [4]; for present purposes a brief encapsulation suffices. Dropping the modal operator from (CC) and (TT) and leaving in place merely a material conditional won't do; that much is clear. After all, there are at present no artifacts to which to instantiate x in order to produce a true antecedent: there is now no computational artifact that can pass the Turing Test. So

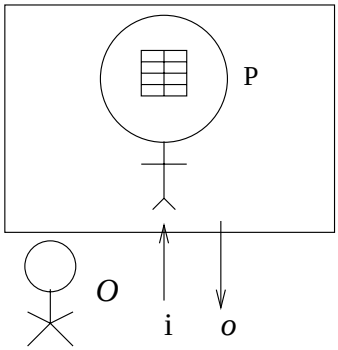

Figure 3: Pictorial Schema of Searle's Response to the Original System Reply, i.e., CR_3.

(CC) Necessarily, if x is a system of suitably configured computation, then x has genuine mental states, which include phenomenal consciousness.

Note that Searle does intend to overthrow a proposition as broad as (CC) via his CRA. He is not just trying to overthrow the view that a system of suitably configured computation would have true linguistic understanding, where this understanding is construed so as not to include phenomenal consciousness. For example, here is Searle summing up the target of CRA in *Minds, Brains, and Science*:

> The point I am making is that if we are talking about having mental states, having a mind, all of these simulations are simply irrelevant. It doesn't matter how good the technology is, or how rapid the calculations made by the computer are. If it really is a computer, its operations have to be defined syntactically, whereas consciousness, thoughts, feelings, emotions, and all the rest of it involve more than syntax. ([20], p. 37)

CRA is also often assumed to target Alan Turing's [23] famous "imitation game" test of computational consciousness, now known as the **Turing Test**, or just 'TT' (without parentheses, to be distinguished from the *proposition* (TT) *about* TT) for short. Accordingly, we specifically assume that CRA also yields the conclusion that the following proposition is false.

(TT) Necessarily, if x passes TT, then x has true understanding of the terms and concepts involved in the conversations that allow x to pass, where 'true understanding' is such that if x has it, then x also has relevant phenomenal consciousness.

So if x is a system of suitably configured computation that is able to engage in conversations about rich, dark, chocolate ice cream, x truly understands such food. Moreover, during the conversation, x will have a number of phenomenal states of consciousness (e.g., remembering that which it's like to savor such ice cream). (It's commonly believed that Turing's goal, in the paper in which he introduced TT [23], was to *eliminate* talk of such things as phenomenal consciousness. But actually, in his response to the Argument from Consciousness, given by Jefferson as an attack on TT, Turing clearly affirms (TT).)

At this point we have

- If CRA is sound, then (CC) is false.
- If CRA is sound, then (TT) is false.

Next, it will greatly facilitate matters if we have on hand a schematic representation of Searle's original thought experiment, the Chinese Room; we call this representation 'CR_1.' (CR_1 is shown diagrammatically in Figure 1.) As we proceed, it will be necessary to modify CR_1 so as to produce variations CR_2, CR_3, ... We will assume that CRA can be suitably adjusted to appeal to a given CR_i in one or more of its premises; we denote the results of these adjustments with a subscript, so that the original $\text{CRA} = \text{CRA}_1$. The symbols i and o refer to the input and output to and from the system S. (In Figures 1, 2, and 3, the system S is composed of the contents of the outermost box or rectangle.) O refers to the outside observers (native Chinese speakers in CR_1), who see i going

the material conditionals would be vacuously true — and all debate would be preempted. What Turing and other computationalists have in mind is that there is some sort of *conceptual* connection between the relevant computation and cognition. This is in general why thought experiments of the right sort can threaten computationalism. The idea behind these thought experiments is that they are cases where the antecedent of a modal conditional is true, but where the consequent isn't.

in and o returning, and ascribe to S genuine understanding of Chinese. P denotes the person who "becomes" a computer in CRA; in Searle's original formulation P = Searle. Finally, P's "rulebook" is shown by the obvious icon.

Given this schematization, it should be wholly uncontroversial that central to all CRA_i are the following locutions.

- S understands i (o).
- O understands i (o).
- P understands i (o).

According to CRA_1, for both i and o, the second is true (and a premise), while the third is false (a premise), as is the first (part of CRA_1's conclusion).

2 The Original Robot and System Reply, and Searle's Responses

The schema developed in the previous section allows us to quickly identify the original System Reply (SR_1) and the original Robot Reply (RR_1) with the following two arguments, respectively.

System Reply$_1$ (SR_1)

(1) If P implements a proper part of S in CR_i, then CRA_i is unsound.
(2) P implements a proper part of S in CR_1.
$\therefore$ (3) CRA_1 is unsound.

(Original) Robot Reply (RR_1)

(4) If in CR_i S doesn't include devices for transforming — by a causal process of the sort that operates when robots interact with the physical world — i into a machine-manipulable encoding $T(i)$, and a machine-manipulable encoding $T(o)$ into o, then CRA_i is unsound.
(5) In CR_1 S doesn't include devices for transforming, by a causal process of the sort that operate when robots interact with the physical world, i into a machine-manipulable encoding $T(i)$, and a machine-manipulable encoding $T(o)$ into o.
$\therefore$ (3) CRA_1 is unsound.

Searle's responses are shown pictorially in Figures 3 and 2, which denote his rebuttals to SR_1 and RR_1, respectively. In the response to the original System Reply, SR_1, Searle simply changes the thought-experiment to CR_3, one in which premise (2) is false; in response to the original Robot Reply, RR_1, Searle simply changes the thought-experiment to CR_2, one in which premise (5) is false. As should be plain, Figure 2 shows the presence of transducers, devices said in (5) to be lacking in S; $T(i)$ and $T(o)$ denote the result of transduction in both directions. Figure 3 shows the result of Searle's having internalized the rulebook. To conclude this section, let's suppose that these thought-experiments are associated with two corresponding variations on CRA_1, viz., CRA_2 and CRA_3.

3 Harnad's Modernized Robot Reply

One of the better known champions of CRA is the psychologist Steven Harnad. Harnad has said on a number of occasions (e.g., [15]) that Searle has overthrown the linguistically-oriented TT by showing that a machine can pass this test by mindlessly moving symbols around in the complete absence of understanding. So Harnad agrees that (TT) and (CC) are overthrown by CRA_2 and CRA_3. How is it, then, that Harnad is to be counted a champion of some version of the Robot Reply? The answer is that Harnad proposes a new target for Searle, one supposedly insulated from CRA_1 through CRA_3. The new target is the *Total* Turing Test (TTT), in which a victorious machine not only displays convincing linguistic behavior, but displays compelling *sensorimotor* behavior as well. In the original Turing Test, you as judge might ask both hidden-from-view contestants to describe the raw sensations they feel upon engaging in their favorite hobby; in the Total Turing Test you might *watch* both contestants (say) play golf, and might then proceed to discuss with them how it feels to smash a picture-perfect drive. When the Total Turing Test is substituted for the Turing Test, Harnad believes that a new sort of robot reply arises. In a symposium in the journal *Think*, in which Searle, Harnad, and Bringsjord participated, Harnad expresses this reply in the form of a dilemma:

> Now back to the Chinese room, but this time TTT-scale rather than just TT-scale. This time, instead of asking whether the TT-passing candidate really understands Chinese or is merely systematically interpretable as if he were understanding it, we will ask whether the TTT-passing candidate (a robot now) really sees the Buddha statue before him or is merely systematically interpretable as if he were seeing it. The robot, in order even to be interpretable as seeing, must have optical transducers. What about Searle, who is attempting to implement the TTT robot without seeing, as he implemented the TT-[passing system] without understanding? There are two possibilities: either Searle receives only the *output* of an optical transducer — in which case it is no wonder that he reports he is not seeing, because he is not implementing the whole system, only part of it, and hence ... the System Reply [= our SR_1] would be correct; or Searle actually looks at the Buddha, in which case he would indeed be implementing the transduction, but then, unfortunately, he *would* be seeing ([16], 17).

This argument conforms to an inference rule known as **constructive dilemma**. According to this rule, if one knows that a proposition p or q is true, and both that if p is true so is r and if q is true so is r, one can conclude r. In Harnad's argument, p is "Searle receives only the output of an optical transducer," and q is "Searle actually looks at the Buddha." According to Harnad, both of these possibilities lead to r, the destruction of CRA. Constructive dilemma is certainly unexceptionable, but let's look more closely at the argument, and let's do so by drawing upon our established scheme. Doing so will quickly reveal a fatal flaw in Harnad's reasoning.

Denote the new thought-experiment here by 'CR_4,' and the associated argument by RR_2. Now, recall the trio of locutions from our schematization of Chinese Room-like arguments:

- S understands i (o).
- O understands i (o).
- P understands i (o).

Recall as well that these locutions are to be instantiated so that the first and third are false (for both i and o), while the second is true (again, for both i and o). Are they instantiated that way in Harnad's new story? They cannot be, for in this story, understanding is supplanted by seeing, and the locutions simply don't refer to seeing. If we try to construe seeing in RR_4 as understanding has

been construed previously, which is surely Harnad's intention, note that seeing here is ambiguous between, as one might say, *merely* seeing and *really* seeing. Suppose we put down before you an object that leaves you utterly clueless; you have no idea whatsoever what this object is. Nonetheless, we might begin our interview with you, after depositing the mystery object on the table before you, by saying: "Okay, now, do you see this thing here?" "Uh, yes," you might well reply — with utter puzzlement. "Well," we might continue, "what do you think it is?" And you might reply that you have no idea. In this case you see the object without understanding; we will call this seeing$_{\bar{u}}$. When you see *with* understanding, we will say that you see$_u$. For Harnad's argument to have a chance of succeeding, he must be read as asking (note the subscript now added to this quote): "What about Searle, who is attempting to implement the TTT robot without seeing$_u$, as he implemented the TT-[passing system] without understanding?" Put in terms of our schemas, this question becomes

- In Harnad's new story, CR_4, is it necessary that P sees$_u$?

Harnad's answer, of course, is an affirmative one — one given, as we've noted, on the strength of the constructive dilemma inference rule. If we concede that one side of the dilemma, the side according to which Searle fails to implement the entire TTT-passing system, does lead to an overthrow of CRA via SR_1, we can focus our attention on the other side, which is expressed, recall, as follows.

> ... or Searle actually looks at the Buddha, in which case he would indeed be implementing the transduction, but then, unfortunately, he *would* be seeing ([16], 17).

The problem should now be evident; it is this. While Searle = P in CR_1, CR_2, and CR_3, fails to understand the objects in question (Chinese inscriptions), such is not the case with respect to Harnad's new thought-experiment. Searle, in "real life," understands what a Buddha is; when he sees one, he sees$_u$ one. It is really no surprise, then, that Harnad gets the result he does! Suppose that some AIniks build a robot able to interact smoothly with — Smuddas. (You don't have a clue what a Smudda is. They look vaguely like salamanders, but are larger, seem somewhat mechanical, and don't appear to be alive in any way.) Now suppose that, in thought, Searle implements the TTT-passing, Smudda-handling system. Would it still be true that Searle sees$_u$ the Smudda? No. Searle will see$_{\bar{u}}$ the Smudda, just as he sees$_{\bar{u}}$ the squiggle squoggles in the original CR. And so we have every reason to believe here that S only sees$_{\bar{u}}$ Smuddas. Harnad's argument fails.

4 The Missing Thought-Experiment

4.1 History of the Missing Thought-Experiment

Harnad's argument marks the *simultaneous deployment of SR_1 and RR_1*; it's from this feature that the argument derives what power it has. The simultaneous deployment of SR_1 and RR_1 is a move that apparently started with Daniel Dennett, in the commentary he offered in the original 1980 *BBS* dialectic [11]; here's what Dennett said there:

> Putting both modifications [= our CR_2 and CR_3] together, we are to imagine our hero controlling both the linguistic and nonlinguistic behavior of a robot who is — himself! When the Chinese words for "Hands up! This is a stickup!" are intoned directly in his ear, he will uncomprehendingly (and at breathtaking speed) hand simulate the program, which leads him to do things (*what* things — is he to order himself in Chinese to stimulate his own motor neurons and then obey the order?) that lead to his handing over *his own* wallet while begging for mercy, in Chinese with his own lips. ([11], 129)

Nowhere in his commentary does Dennett give an argument against CRA based on combining SR_1 and RR_1; he only calls for a thought-experiment that would seem to be missing in the literature to this day. In Dennett's own words again:

> In point of fact, Searle has simply not told us how he intends us to imagine the case, which we are licensed to do by his two modifications [= our CR_2 and CR_3] ... There are several radically different alternatives — all so outlandishly unrealizable as to caution us not to trust our gut reactions about them in any case. When we imagine our hero "incorporating the entire system" are we to imagine that he pushes buttons with with his fingers in order to get his own arms to move? Surely not, since all the buttons are now internal. Are we to imagine that when he responds to the Chinese for "pass the salt, please" by getting his hand to grasp the salt and move it in a certain direction, he doesn't *notice* that this is what he is doing? In short, could anyone who became accomplished in this imagined exercise fail to become fluent in Chinese in the process? Perhaps, but it all depends on details of this, the only crucial thought experiment in Searle's kit, that Searle does not provide. ([11], 129)

There are in this quote the *seeds* for an argument that CR_1–CR_3 are actually *not* conceivable. has Dennett elsewhere cultivated these seeds into a full-blown argument for the view that that CR_2 and CR_3 are conceivable only if Searle-in-them understands Chinese? No, not to our knowledge. Dennett *has* argued elsewhere that plain old CR_1 is conceivable only if Searle-in-CR_1 understands Chinese. Dennett gave this argument 11 years after the original *BBS* exchange in his *Consciousness Explained* [11]. After he offers a sample exchange between a judge and a computer in the Turing Test, he argues as follows:

> The fact is that any program that could actually hold up its end in the conversation depicted would have to be an extraordinarily supple, sophisticated, multilayered system, brimming with "world knowledge" and meta-knowledge and meta-meta-knowledge about its own responses, the likely responses of its interlocutor, its own "motivations" and the motivations of its interlocutor, and much, much more. Searle does not deny that the program can have all this structure, of course, He simply discourages us from attending to it. But if we are to do a good job imagining the case, we are not only entitled but obliged to imagine that the program Searle is hand-simulating has all this structure — and more, if only we can imagine it. But then it is no longer *obvious*, I trust, that there is no genuine understanding of the joke going on. ([14], 438)

This is really an astonishingly bad argument. It's an elementary mathematical fact that all computer programs can be recast as exactly equivalent Turing machines operating with only a binary alphabet such as {0, 1}. Instructions here consist in nothing more than such imperatives as "If the machine is in internal state n, with its read/write head scanning a 0, erase the 0 and write a 1, and then have the machine enter internal state m." (Such instructions are often given as austere quadruples; in this case the quadruple would be $(n01m)$.) So in response to Dennett, let's stipulate that Searle-in-CR_1 manipulates only such instructions. It should be obvious that if all Searle is doing is hand-simulating in this way, he will have no understanding of Chinese, no understanding of what the original high-level program is about.[4] But what about CR_2 and CR_3? Can their simultaneous deployment, that is, the missing thought-experiment, be specified? If not, Dennett will nonetheless have managed to triumph over Searle.

4.2 Searle's Steps Toward the Missing Thought-Experiment

Searle himself, in reply to Harnad's modernized robot reply, can be read as trying to describe, or at least move toward a description of, the missing thought-experiment. Here's what Searle says:

[4] Of course, there remains the issue of *speed*: in order to carry out such hand simulation, Searle is going to have to work rapidly. But as one of us has explained elsewhere [5], speed is a red herring.

> Harnad thinks it is an answer to this to suppose that the transducers have to be part of me and not just an appendage. Unless they are part of me, he says, I am "not implementing the whole system, only part of it." Well fine, let us suppose that I am totally blind because of damage to my visual cortex, but my photoreceptor cells work perfectly as transducers. Then let the robot use my photoreceptors as transducers in the above example [= CR_2]. What difference does it make? None at all as far as getting the causal powers of the brain to produce vision. I would still be blindly producing the input output functions of vision without seeing anything. ([22], 69-70)

As we've indicated by our parenthetical in this quote, by "the above example" Searle is referring to CR_2, the scenario he offers as a rejoinder to the original Robot Reply, RR_1.[5] But does the introduction of blindness and the robot's use of Searle's photoreceptor cells, added to CR_2, constitute the missing thought-experiment? No, for there are two problems.

The first problem is that in this new scenario Searle is simply no longer doing mindless symbol manipulation! If he is blind, and the robot is using his photoreceptors, then optical stimuli is converted to symbolic data, whereupon this data is manipulated — but it isn't *Searle* who is carrying out the manipulation. To see this, consider the following scenario. While you're sound asleep, Dr. Black, a skilled but amoral neurosurgeon, sneaks into your bedroom and anesthetizes you. Then he proceeds to surgically appropriate your visual system so that it is attached to a computer-based voice production system. Black is thus able to hold up objects before your eyes and a synthetic voice announces what the object is. So if Black holds up an apple, a synthetic voice says, "Apple." Now in this case symbols are being manipulated by your visual system (and other biological processes that may or may not at bottom by symbol manipulation in action). But *you* are not manipulating symbols. Likewise, while *Searle* is manipulating squiggle squoggles in the original CR, his photoreceptor cells, not him, are manipulating symbols in his modification of CR_2.

The second problem is by now a familiar one: in his blindness story Searle is something well short of the entire system, so the original System Reply kicks in against him. His photoreceptors have become part of the overall system, yes, but lots of the system is composed of non-Searlean stuff.

We offer a way for Searle to surmount the first problem; here's our suggestion. Add to the thought-experiment that Searle manipulates, in accordance with the relevant computer program, the *braille-based* representation of the symbolic information emerging from his photoreceptors. This modification injects back into the thought-experiment the fact that *Searle* is carrying out purely syntactic symbol manipulation.

[5] The scenario is redescribed in the *Think* symposium as follows.

> Imagine a really big robot whose brain consists of a commercial computer located in a Chinese room in the robot's cranium. Now replace the commercial computer with me, Searle. I am now the robot's brain carrying out the steps in the robot's programs. The robot has all of the sensory and motor transducers it needs to coordinate its input with its output. And I, Searle, in the Chinese room am doing the coordinating, but I know nothing of this. For example, among the robot's transducers, we will suppose, are devices that convert optical stimuli into Chinese symbols. These Chinese symbols are input to the Chinese Room and I operate on these symbols according to a set of rules, the program. I operate on these symbols and eventually send symbols to transducers that cause motor output. The motor output is an utterance that says in Chinese, "I just saw a picture of a big fat Buddha." But all the same, I didn't see anything, and neither did the robot. That is, there were no conscious visual experiences of a Buddha in question. What actually occurred was that light sensitive detectors in the robot's skull took in stimuli, converted these into symbols, I processed the symbols and sent them to an output transducer that converted the symbolic output into an auditory utterance. In such a case you can have all of the transducers you want and pass the TTT until the sun goes down but you still do not thereby guarantee the appropriate experience, understanding, etc. ([22], 69)

What about the second problem? This one isn't so easily solved. Searle himself does see the problem, and his response to Harnad is interesting, but it's almost as if Searle gives this response half-heartedly, sure in his heart that his general attack is ultimately victorious, but a bit tired with the dialectic — too tired to continue to engage Harnad point for point and slug things out till the final bell sounds. Here is what Searle says:

> Will Harnad insist in the face of this point that I am still not implementing the whole system? If he does then the thesis threatens to become trivial. Any system identical with me has cognition because of my neurobiological constitution. Anything identical with my system can cause what my system causes, and all the talk about TTT, computation and the rest of it would now become irrelevant. ([22], 69)

Searle is of course correct that P in CR_i will invariably be instantiated to something that "has cognition." But the trick is instantiating P in such a way that $P = S$, where P fails to understand what O believes S does understand. Searle pulls out of the search for a scenario that includes these facts.

However, the second problem *can* be surmounted; the missing thought-experiment *can* be specified; Dennett's challenge can thereby be met. How? Searle is on the right track: imaginative surgery that serves to blur the distinction between robots and persons offers the key to providing the missing thought-experiment. We frame the description of such surgery by discussing a related "surgical" project in our Minds *&* Machines Laboratory. We put 'surgical' in scare quotes here because the process in question is not literally a surgical one. The process is aimed at producing "zombie animals,' or just "zombanimals," qualia-less computational simulacra for simple biological animals. It involves abstracting the biological so as to produce a model (or, as we prefer to say, following Pollock [18], a *blueprint*) that can then be rendered computational. Because the focus in the present paper is thought-experimental, we will now illustrate this process by describing it in connection with a series of *imaginary* biological creatures, rather than with real ones. Doing so will allow us to avoid biological and engineering details that are irrelevant to the arguments at stake in this paper.

4.3 Framing the Missing Thought Experiment with Simple Zombanimals

Most of you are familiar with the thought-experiments in which zombies — creatures displaying our external behavior but bereft of consciousness — are described; here's one in brief. You're diagnosed with inoperable brain cancer that will inevitably and quickly metastasize. Desperate, you implore a team of neurosurgeons to replace your brain, piecemeal, with silicon chip workalikes, until there is only silicon inside your refurbished cranium.[6] The procedure is initiated ... and the story then continues in a manner that seems to imply the logical possibility zombies. In (none other than) Searle's words:

> As the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior ... [You have become blind, but] you hear your voice saying in a way that is completely out of your control, 'I see a red object in front of me.' ... We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same ([21], 66-7).

[6] The silicon supplantation is elegantly described in [8].

Lots of people — Searle, Dennett, Chalmers, Bringsjord, etc.[7] — have written about zombies. Most of these thinkers agree that such creatures are logically possible, but, with the exception of Bringsjord, they hold that zombies are *physically im*possible. That is, most refuse to accept that such surgery can *in fact* be carried out. In order to frame the missing thought-experiment, we begin by trying to indicate that such surgery can indeed be carried out — *for animals.*

Begin by imagining a very simple animal. Not a cat or a rat; they're too sophisticated. Imagine a simple multi-cellular organism; let's call it a 'bloog.'[8] When you shine a penlight on a bloog, it propels itself energetically forward. If you follow the bloog as it moves, keeping the penlight on it, it continues ahead rather quickly. If you shut the penlight off, the bloog still moves ahead, but very, very slowly: the bloog is — we say — listless. If you keep the penlight on, but shine it a few inches away and in front of the bloog ('front of' being identified with the direction in which it's moving), the bloog gradually accelerates in a straight line in the direction it appears to be facing, but then slows down once it is beyond the light.

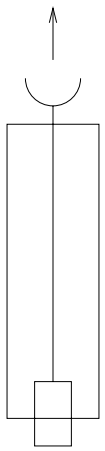

Figure 4: The Zombanimal Robot V1. *The motor is denoted by the rectangular box at the tail end, the sensor by the half-circle on a stalk.*

A roboticist in our Minds *&* Machines Laboratory, Clarke Caporale, spent some time experimenting with a bloog for a while with his penlight, and witnessed the behavior we have just described; he then set to work.[9] Clarke began by scanning the flow of information in a bloog when one is under a microscope. After a bit, he readied his supply of robotics micro-hardware, and began to operate. Now that he is done, he presents you with ... a creature he calls 'V1.'[10] V1 is composed of one tiny sensor and one tiny motor, which are connected, and a biological structure

[7]For Searle, see [21]. For Dennett, see [13], [12], and [14]. For Chalmers, see [7]. Bringsjord's main contribution, again, is in "The Zombie Attack on the Computational Conception of Mind" [4], which builds on Searle's [21] take on the thought-experiment just given.

[8]The word 'bloog,' as well as 'sneelock,' 'fleelock,' 'moog,' and 'multi-moog,' words used below to name other organisms, have no special linguistic significance, but are used here in deference to sorts of names used by the inimitable (and, alas, late) Dr. Seuss, who received a PhD from Brown University the same year Selmer did.

[9]Remember that though bloogs and their relatives (soon to be described) are imaginary, all the robotics is real. The biology is dropped in this paper to ease exposition. There is a photograph of the actual robotics workbench used by Caporale at

`http://www.rpi.edu/~bestlj/SELPAP/SEARLEBOOK/workbench.jpg`

[10]Our simple zombanimals are inspired by the vehicular creatures described in Valentino Braitenberg's *Vehicles: Experiments in Synthetic Psychology* [1]. Our first zombanimal is Braitenberg's Vehicle 1, or just 'V1' for short.

left over from the original bloog that supports them. The motor is connected to some device which, when driven by the motor, produces locomotion. V1 is shown in Figure 4. The behavior of V1 is straightforward: the more of the source detected by its sensor, the faster its motor runs. You will have to take our word for it: though, as we've said, the biological creatures are imaginary, V1 is real, and really behaves as described.[11] If V1 is bathed in light from the penlight, it moves forward energetically. If it then enters a darker area it becomes listless. If it detects a light ahead, it accelerates toward the light, passes through it, and then decelerates.

Now consider two more complex biological creatures, a 'sneelock' and a 'fleelock.' They are larger than bloogs, a slightly different shade of fleshy brown, and behave differently. A fleelock behaves as follows. If you shine a penlight on the surface on which the fleelock is located, just ahead and exactly in front of the organism, it moves directly toward the light and passes through it, at which point, like bloogs, it becomes listless. However, if you shine the penlight ahead of the fleelock, but to the left or right, it turns to avoid the light, and then moves forward slowly; fleelocks generally dislike light. Sneelocks are similar. They too dislike light, but there is a difference: sneelocks are aggressive. This can be shown by shining a penlight ahead of a sneelock (and, again, to the left or right). When one does this, the sneelock turns with increasing rapidity toward the light, and moves directly at it, eventually moving frontally into the light to apparently assault it.

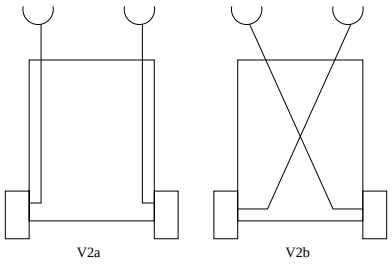

Figure 5: V2a and V2b. *V2a orients away from the light; V2b toward it.*

Caporale performed surgery once again. The result is a pair of new zombanimals, V2a and V2b (see Figure 5). Courtesy of micro-sensors and motors, V2a behaves just like a fleelock, V2b just like a sneelock. V1, V2a, and V2b behave just as their biological counterparts do.

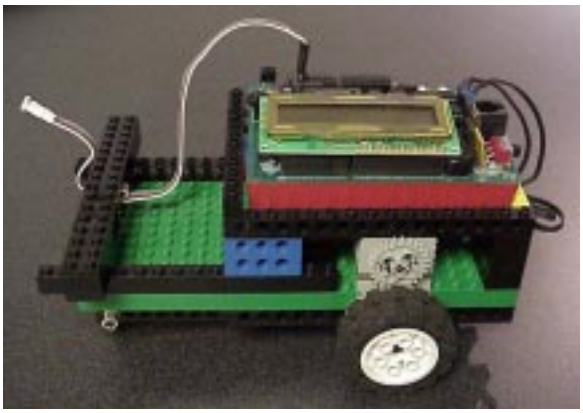

Figure 6: Photo of Caporale's Actual V1

[11] At the 1998 Eastern Meeting of the American Philosophical Association, for a talk entitled "Zombanimals," given by the two of us, Clarke provided a working physical demonstration of V1, V2a, V2b, and V3. The talk was given to the Society for Machines and Mentality.

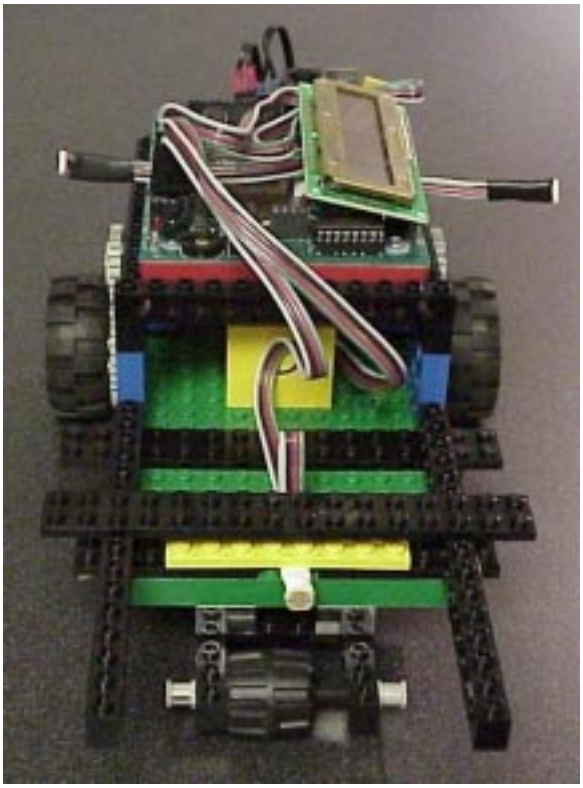

Figure 7: Photo of Caporale's Actual V2a

You're doubtless thinking that such organisms as bloogs, sneelocks, and fleelocks are excruciatingly simple. Well, you're right. As we've indicated, they're *simple* zombanimals. But Caporale is just warming up.

Consider an animal that can sense and react to not only light, but temperature, oxygen concentration, and amount of organic matter. This biological creature is called a 'multi-moog.' Multi-moogs dislike high temperature, turn away from hot places, dislike light with considerable passion (since it turns toward and apparently attempts to destroy them), and prefers a well-oxygenated environment containing many organic molecules. Caporale has "zombified" a multi-moog; the result is V3c, shown in Figure 8. V3c has four pairs of sensors tuned to light, temperature, oxygen concentration, and amount of organic matter. The first pair of sensors is connected to the micromotors with uncrossed excitatory connections, the second pair with crossed excitatory connections, and the third and fourth pairs with inhibitory connections.

4.4 Specifying the Missing Thought Experiment Courtesy of Real Robots

We are now in position to articulate the missing thought-experiment. Suppose, first, that AI has progressed to the point of producing a robot sophisticated enough to pass Harnad's TTT — a robot of the sort that Pollock and Moravec and Hayes . . . predict will forever silence Searle's CRA; call this robot 'R.' Accordingly, we will have on hand a blueprint B_R for R that is generally like the blueprints for the robots V1, V2a, V2b, and V3 given in the corresponding figures 4, 5, and 8, except that we stipulate that B_R includes the program P_R for R that makes "executive" decisions about the flow of information to and from R's brain A_R. (Needless to say, B_R will be *very* complex — but Pollock [18] has already offered a good start.) Now we introduce Searle as the star of yet another gedanken-experiment. We are going to gradually use him to implement parts

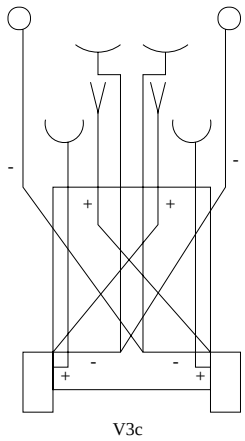

Figure 8: V3c. *A multisensorial zombanimal.*

and processes of R by studying B_R and carrying out the required surgery. (We are thus simply reversing the procedure discussed in connection with bloogs, sneelocks, fleelocks, and multi-moogs.) Let's start by giving Searle drugs so that he no longer has a sense of touch or proprioception. Now we proceed to tackle locomotion. Without loss of generality, we can assume that there is an effector E_R in R responsible for R's ability to walk, and also that E_R is activated when the "brain" A_R of R, as controlled by P_R, sends the appropriate signals. So let's perform some surgery on Searle's legs and feet: let's activate his feet for walking by the signals from R's brain A_R. If we have a way of piping to Searle's legs and feet signals of the sort that routinely flow from A_R, we have a way of causing Searle to walk entirely beyond his control. And we can pull off precisely the same trick for R's other effectors. For example, whenever R waves its robotic hand, this will happen because of signals passing from A_R to the relevant effector. So we can appropriate Searle's arms and hands, and test them out by sending to them the relevant signals. This will allow us to cause Searle to wave for reasons, again, beyond his control. Next, as our readers will by this time have guessed, we focus on sensing. Let S_R be a visual sensor for R. Imagine that we blind Searle in such a way that his photoreceptor cells continue to — as he put it above in the thought-experiment that started the move toward to the one we're presenting here — "work perfectly as transducers." Next, let Searle's photoreceptors transduce stimuli sent to S_R and pipe the resulting data to a console that allows us to view this data.

Why a console? We refer to one in order to sharpen your mental picture of what we have so far. Imagine the surgery we've described taking place in an elaborate operating room in which lies the "wired up" Searle. Picture as well that you have before you the console, and that based on what data you see on its display, you can mindlessly obey P_R in order to push buttons to send signals to the parts of Searle's body that have been taken over. We have reached the point where you can imagine this: you put an object in front of Searle's eyes, a Buddha, b, say, and you observe the data that results from Searle seeing$_{\bar{u}}$ b, and then, on the basis of what you see, you send signals to Searle's arm and hand in order to wave his hand. At this point the situation is summed pictorially in Figure 9.

Obviously, the idea is that you now continue to painstakingly use study of the blueprint B_R and the program P_R, along with surgery, to turn Searle's body into the perfect biological counterpart for all of R's input/output behaviors. The final step is supplanting you, the console, and P_R with something that stays internal to Searle. In order to pull off this step, we bring in elements of the "CR-ish" thought-experiment described by Bringsjord in the chapter "Searle" from his *What*

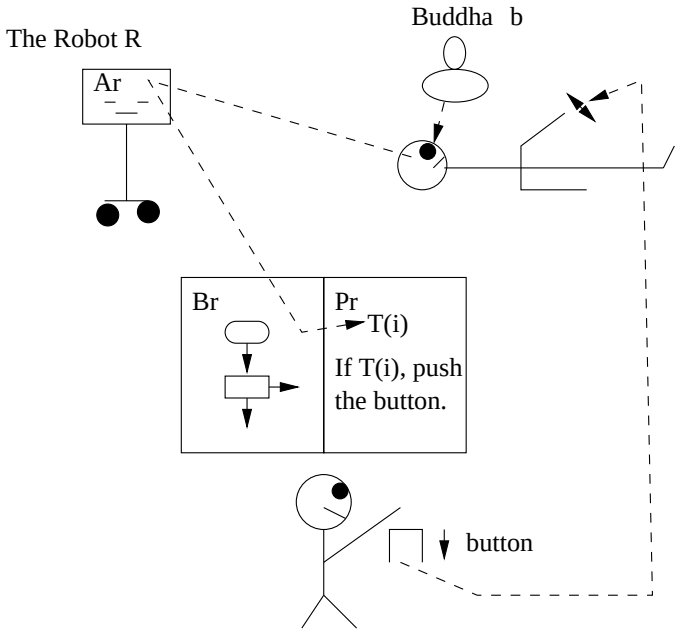

Figure 9: The Situation As We Approach the Missing Thought-Experiment

Robots Can and Can't Be [2]: we imagine that Searle is now wired so that *what you see on your console he sees in the form of mental images.* We also imagine that by manipulating certain mental images in certain ways, effectors in Searle's body are activated. For example, Searle causes his hand to wave (without, of course, his sensing that it has moved) by visualizing the mental image of a button being depressed.) Suppose, then, that, on the basis of what he sees with his mind's eye, Searle can follow P_R to route signals to the parts of him that have been wired up in parallel to B_R. At this point, when his incisions heal up and he heads out of the operating room into the world, Searle will convince all outside observers that he (say) sees_u a Buddha and that he understands that some Chinese speaker just asked him if he likes hamburgers. But clearly Searle sees_u absolutely nothing. Searle thus embodies a decisive rebuttal to the combination system/robot reply: the missing thought-experiment has been found. (Figure 10 shows, schematically, the missing thought-experiment.)

References

[1] V. Braitenberg. *Vehicles: Experiments in Synthetic Psychology.* Bradford Books, Cambridge, MA, 1984.

[2] S. Bringsjord. *What Robots Can and Can't Be.* Kluwer, Dordrecht, The Netherlands, 1992.

[3] S. Bringsjord. Could, how could we tell if, and why should–androids have inner lives? In K. Ford, C. Glymour, and P. Hayes, editors, *Android Epistemology*, pages 93–122. MIT Press, Cambridge, MA, 1995.

[4] S. Bringsjord. The zombie attack on the computational conception of mind. *Philosophy and Phenomenological Research*, 59.1:41–69, 1999.

[5] S. Bringsjord and D. Ferrucci. Reply to thayse and glymour on logic and artificial intelligence. *Minds and Machines*, 8:313–315, 1995.

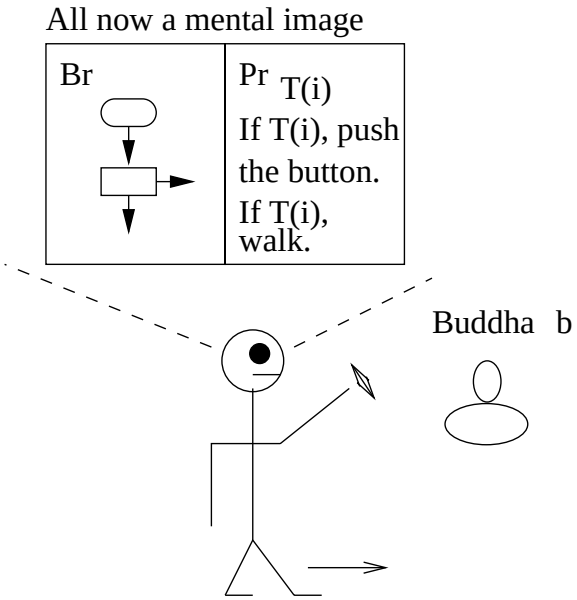

Figure 10: The Missing Thought-Experiment

[6] S. Bringsjord and M. Zenzen. Cognition is not computation: The argument from irreversibility? *Synthese*, 113:285–320, 1997.

[7] D. Chalmers. *The Conscious Mind: In Search of a Fundmental Theory*. Oxford, Oxford, UK, 1996.

[8] D. Cole and R. Foelber. Contingent materialism. *Pacific Philosophical Quarterly*, 65(1):74–85, 1984.

[9] W. Davis. *The Serpent and the Rainbow*. Simon & Shuster, New York, NY, 1985.

[10] W. Davis. *Passage of Darkness: The Ethnobiology of the Haitian Zombie*. University of North Caronlina Press, Chapel Hill, NC, 1988.

[11] D. Dennett. The milk of human intentionality. *Behavioral and Brain Sciences*, 3:128–130, 1980.

[12] D. Dennett. Review of Searle's *the rediscovery of the mind*. *Journal of Philosophy*, 90(4):193–205, 1993.

[13] D. Dennett. The unimagined preposterousness of zombies. *Journal of Consciousness Studies*, 2(4):322–326, 1995.

[14] D.C. Dennett. *Consciousness Explained*. Little, Brown, Boston, MA, 1991.

[15] S. Harnad. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1.1:43–54, 1991.

[16] S. Harnad. Grounding symbols in the analog world with neural nets: A hybrid model. *Think*, pages 12–20, 1993.

[17] H. Moravec. *Robot: Mere Machine to Transcendant Mind*. Oxford University Press, Oxford, UK, 1999.

[18] J. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA, 1995.

[19] J. Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.

[20] J. Searle. *Minds, Brains, and Science*. Harvard University Press, Cambridge, MA, 1984.

[21] J. Searle. *The Rediscovery of the Mind*. MIT Press, Cambridge, MA, 1992.

[22] J. Searle. The failures of computationalism. *Think*, pages 68–71, 1993.

[23] A. Turing. Computing machinery and intelligence. In A. R. Anderson, editor, *Minds and Machines*, pages 4–30. Prentice-Hall, Englewood Cliffs, NJ, 1964.