CDD: 001.535

ILLUMINATING THE CHINESE ROOM

TERRY DARTNALL

Computing and Information Technology,
Griffith University,
Brisbane,
AUSTRALIA 4116

TERRYD@CIT.GU.EDU.AU

In this paper I provide a solution to the problem of the Chinese Room. The problem is to determine whether the Chinese Room Argument goes through, and if it does, to explain why symbol handling does not give us cognition. I argue that the real issue is not about symbols, but about the relationship between cognition and content. Artificial Intelligence (AI) does not distinguish between these, and naively believes that internalising the public symbolisms that express the content of cognition will generate cognition itself. Not only does it do this in practise: the main manifestos of AI explicitly state that the internalised symbolisms are interpretted and contentful.

This confusion between cognition and content is the same confusion that underlies psychologism, which says that we can find out about content by studying cognition. What I call "reverse psychologism" says that we can find out about cognition by studying content, and in its stronger form, that we can generate cognition by internalising content. This is the real fallacy that is exposed by the Chinese Room Argument.

© *Manuscrito*, 1996. Published by the Center for Logic, Epistemology and History of Science (CLE/UNICAMP), State University of Campinas, P.O. Box 6133, 13081-970 Campinas S.P., Brazil.

INTRODUCTION

In this paper I provide a solution to the problem of the Chinese Room. I do this by providing an argument that, like Searle's, shows that symbol manipulation cannot generate cognition. Strong AI, as eponymised in the Physical Symbol System Hypothesis and the Knowledge Representation Hypothesis, cannot deliver the goods.

But why can't it deliver the goods? Even those who are persuaded by Searle's argument are frustrated by its opacity. *Why* is there no understanding in the room, and *why*, more generally, is symbol handling unable to give us cognition?

I argue that strong AI confuses cognitive states with the *content* of those states, and tries to generate cognition by locating contentful symbol structures inside a system. My emphasis is therefore on the distinction between cognition and content, rather than the relationship between cognition and symbols (though the former throws light on the latter). It is true that classical, symbol handling AI manipulates formally specified elements according to formal (i.e. syntactic) rules. Nevertheless, these elements, and the structures that they constitute, are contentful to us, and that is why classical AI internalises them. It tries to generate cognition by internalising content, and it tries to do this by internalising symbols that express that content. Thus it is the relationship between cognition and content that is really the issue. Content is internalised, and the symbolic vehicle comes along for the ride.

The distinction between cognition and content is part of a more general distinction between what is cognitive and what is not, and it applies to connectionist systems as much as classical ones. Assigning content to nodes or patterns of activation

© *Manuscrito*, 1996.

in a trained-up network is another case of trying to get cognition by internalising content.

The confusion between cognition and content has a curious history. In the nineteenth century it gave rise to psychologism, which is the belief that we can study disciplines such as logic and mathematics by studying the mind, so that these disciplines are branches of empirical psychology. The official story is that psychologism was exorcised by Frege and Husserl and buried at the crossroads of intellectual history. Be that as it may, the confusion that underlies it – the confusion between cognition and content – lives on in what I call "reverse psychologism". This is the belief that we can study cognition by studying content, and, in its stronger form, that we can *generate* cognition by internalising content.

The paper comes in three parts. In the first I outline the Chinese Room Argument (hereafter CRA) and the position it is attacking. In the second I examine the content-cognition confusion and show how it led to psychologism in the nineteenth century, and I look at the factors that compound it. I examine reverse psychologism, and show how it arises from the same confusion and is compounded by the same factors. I provide worked examples from linguistic theory and AI. In part three I apply these results to the CRA.

1. THE CHINESE ROOM ARGUMENT

Everyone knows about the Chinese Room, but here it is again.

Searle is seated at a mahogany desk with a nice inlaid leather top. On the desk are pens, pencils, a desk lamp and a cup of coffee with three lumps of sugar. In front of the desk are two windows. Pieces of paper covered in squiggles are

plopping in through one of the windows. Searle examines the squiggles and looks them up in a rulebook (which is next to his cup of coffee). The rulebook is in English, and it tells Searle what to do with the squiggles: he can reproduce them, modify them, destroy them, and/or create new ones, and sometimes he passes the results back through the other window.[1]

Now unbeknownst to Searle, there are Chinese computer programmers outside the room, feeding Chinese sentences into it, and, from their point of view, getting Chinese sentences back in reply. The rule book is so sophisticated, and Searle so adept at using it, that the room appears to understand Chinese, and this is certainly what the programmers believe. But, says Searle, the room understands nothing, for he does not understand Chinese, nor does anything else in the room, and nor do the room and its contents as a whole.

From this, he says, it follows that computers do not understand their input, for they too manipulate input squiggles according to formal rules, or as he puts it, "perform computational operations on formally specified elements".

This notion of performing computational operations on formally specified elements is the heart of the matter, and I have generalised the argument to make this clear. In fact Searle does not cite anyone who specifically makes this claim, but focuses on the work of Roger Schank and his colleagues at Yale (cf. Schank & Abelson (1977)). Schank and Abelson programmed a computer with a script that provides a framework for what we expect when we go into a restaurant: we expect ta-

[1]Searle does not explicitly list these operations. I have borrowed them from Schank & Abelson (1977). I am sure he would endorse them.

© *Manuscrito*, 1996.

bles and chairs, for instance, and not an ocean. Then they told the computer a story, such as: "John went into a restaurant and ordered a hamburger. When it arrived it was burnt to a crisp. John stormed out." They asked the computer "Did John eat the hamburger?", and it answered "No". Searle concedes that this is an interesting result, because the computer hasn't been given this information, but he denies that it understands, for it merely manipulates sets of symbols ("the script", "the story" and "the question") according to formal rules.

Well, is there anyone who specifically claims that "computational operations on formally specified elements" can generate cognition? Yes, this claim is explicitly made by the two hypotheses that underlie classical, symbol-handling AI: Newell and Simon's Physical Symbol System Hypothesis (Newell & Simon (1976)) and Brian Cantwell Smith's Knowledge Representation Hypothesis (Smith (1985)).

The Physical Symbol System Hypothesis says that "A physical symbol system has the necessary and sufficient means for general intelligent action." Newell and Simon go on to say "By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be further organized to exhibit general intelligence."

The Knowledge Representation Hypothesis is more explicit and says that a system knows that p if and only if it contains a symbol structure that means p to us and that causes the system to behave in appropriate ways. Thus, for instance, a system knows that tigers bite if and only if it contains a structure that means "Tigers bite" to us, and that causes it to climb trees in the presence of tigers. In keeping with this belief, knowledge engineers put knowledge structures and belief structures

(frames, semantic networks, production systems, sentences, logic, etc.) into Belief Bins and Knowledge Bins, in the belief that this will give the systems knowledge and belief. Dennett calls this "High Church Computationalism".

We find a similar claim in cognitive science. Fodor's Language of Thought Hypothesis was the only game in town until the re-emergence of connectionism in the mid-80s, and it claims that an essential aspect of cognition is the manipulation of symbols in an innate, inner language that Fodor calls "Mentalese" (e.g. Fodor (1975)).

So the matter seems cut and dried. Searle says he is attacking the claim that computational operations on formally specified elements can generate cognition, and we have found such a claim at the heart of classical AI. But now that we know where to look, we will find that this claim is driven by a deeper confusion, between cognition and content. I part two I examine this distinction in detail. In part three I show how it underlies the symbol handling hypothesis.

2. CONFUSING COGNITION AND CONTENT

The distinction between cognition and content was first drawn by Frege, who said "A proposition may be thought, and again it may be true; never confuse these things." He added "We must remind ourselves, it seems, that a proposition no more ceases to be true when I cease to think of it than the sun ceases to exist when I shut my eyes" (Frege (1967)). Husserl examined the distinction in more detail, and called it the "act/content distinction", although he also characterised it as the distinction between our consciousness of the objects of thought and the objects themselves (Husserl (1962)). I shall call it the "state/content distinction", which I think is clearer.

© *Manuscrito*, 1996.

The distinction we are concerned with is the distinction between a psychological state, such as believing that the earth is round, and the content of that state, which can be expressed in a public, communicable symbolism. In one sense the belief that the earth is round is a cognitive state, but in another it is a proposition that can be written down and that expresses not only the content of *my* belief, but (I assume) the content of yours as well.

In fact all mentalistic terms are ambiguous between their cognitive and non-cognitive senses. Some, such as "belief" and "thought", are ambiguous between state and content, whereas others, such as "love" and "desire", are ambiguous between state and object. "My love is unrequited and wears thigh high boots" equivocates between my state, which is unrequited, and the object of my state, who has nice footwear. "Knowledge" is ambiguous both ways, as we shall see.

We can bring the state/content distinction into relief in two ways. The first is by looking at predicates that contents can take and that states cannot, and, similarly, at predicates that *states* can takes that *contents* cannot. Consider the case of belief. A belief in the sense of content (or proposition) can be true or false, tautologous or contradictory, subscribed to by one or many. It can be written down. Here is a belief in this sense: "Brenda has nice footwear", or, for those who prefer a different sort of example, "snow is white". There is nothing cognitive about beliefs in this sense. On the other hand, beliefs as cognitive states can be strong and passionate, sincere or insincere, shortlived or longlasting, but not true or false, or tautologous or contradictory.

If we confuse these senses we will end up saying that a belief is sincere and tautologous, or that it is contradictory and four years old. We will confuse state and content.

© *Manuscrito*, 1996.

The other way of distinguishing between state and content is to observe that different states can have the same content. We can believe and fear the same thing – that there is no beer in the fridge, for instance, or that Brenda's high heels will make holes in the carpet.

This apparently trivial mistake can lead to fundamental confusions about the conceptual foundations of disciplines. It can lead to psychologism, which is the belief that we can find out about content by examining states, and it can lead to what I call "reverse psychologism", which is the belief that we can find out about states by examining content, or (in its stronger form) that we can *generate* states by internalising content.

Psychologism is well known, but its underlying cause, the state/content confusion, is not. I will look at psychologism in logic and mathematics and show that its cause is the state/content confusion.

2.1. PSYCHOLOGISM IN LOGIC AND MATHEMATICS

The best known advocate of psychologism in logic and mathematics is John Stuart Mill. Mill believed that introspection is the only basis of the principles of logic and the axioms of mathematics (Mill (1843)) and he classified logic under psychology (Mill (1865)). He maintained that logic is the study of consistency relationships between psychological states. Thus the law of non-contradiction (that a proposition cannot be both true and false) is the claim that anyone who is in the belief-state characterised by believing A will not also be in the belief-state characterised by believing not-A. Similarly, the law of the excluded middle (that a proposition is either true or false) is "a generalisation of the universal experience that ... if con-

© *Manuscrito*, 1996.

sciousness is not in one of the two modes it must be in the other". (Mill (1843), bk. 2, ch. 7, sec. 5.)

A list of infelicities that can be laid at the feet of this position. They were first voiced by Frege, and then articulated more thoroughly by Husserl[2].

The contingency argument. If the laws of logic are empirical then they are contingent. If they are contingent then they can be false. But to say, for instance, that the law of non-contradiction can be false is to say "possibly not: ~(A & ~A)", and this is a contradiction.

The contingently false argument. If the laws of logic are empirical generalisations about how we think, then they are not only contingent, but contingently *false*, since some of us are inconsistent some of the time.

The a posteriori argument. If the laws of logic are empirical generalisations then logic would proceed *a posteriori*: we would need to look in the world to discover and test such laws. But we do not do empirical surveys to determine the truth of laws such as the law of non-contradiction[3].

The existential argument. The laws of logic are not about anything in the empirical world and are therefore not about mental states. The law of non-contradiction, for instance, does

---

[2]Husserl's first book, *The Philosophy of Arithmetic*, attempted to base arithmetic on psychological foundations. Frege reviewed it and criticised its psychologism. Husserl acknowledged this criticism and spent the following years studying psychologism in logic and mathematics and formulating arguments against it.

[3]Husserl said 'No natural laws can be known *a priori* ... The only way in which a natural law can be established and justified, is by induction from the singular facts of experience ... Nothing, however, seems plainer than that the laws 'of pure logic' all have *a priori* validity'. Husserl (1970), p. 99.

© *Manuscrito*, 1996.

not quantify over mental states. As Husserl put it, "No logical law implies a "matter of fact""[4].

This looks like a fairly damning list of indictments. In fact the proponent of psychologism can say that the contingency argument begs the question, and that the contingently false argument ignores the role of idealisation in science. I do not think that these moves succeed, but I will not go into the details here. Instead I will rely on the *a posteriori* and existential arguments. We do not establish theorems in logic by doing empirical surveys, and these theorems are not about anything in the world. Psychologism in logic has the wrong ontology and the wrong methodology: logic is not about cognitive states and is not based on empirical investigation.

So exit psychologism in logic. Exit, too, psychologism in general, for the falsity of psychologism in logic demonstrates the falsity of psychologism in general.

Why is this? The point is that the consequences of confusing state and content are *most obvious* in the case of logic. Frege criticised psychologism because he believed that it leads to a kind of "consensus theory of truth": we find out whether it is cold at the North Pole by doing a survey of North-Pole-belief-states. He felt that such a consensus lacked "objective certainty", and he believed that this was most obviously true in the case of logic.

But the case against psychologism is stronger than this. Cognitive states, such as "believing it is cold at the North Pole", are not the sorts of things that can be true or false *at all.* In or-

[4]Husserl said '[If] the laws of logic have their epistemological source in psychological matters of fact [then] ... they must themselves be psychological in content, both being laws for mental states and also be presupposing or implying the existence of such states'. Husserl (1970), p. 104.

der to reintroduce truth and falsity, psychologism has to ascend a level and generalise *about* cognitive states. It has to say that the truth of the sentence "It is cold at the North Pole" consists of everyone being in the state "believing it is cold at the North Pole".

This ascension of levels fails most obviously in the case of logic, where more is lost than truth and falsity. Necessary truth and falsity are lost as well. Mill bit the bullet here and accepted that the laws of logic are contingent. But *a priority* is lost as well: the ascension of levels requires us to say that the laws of logic are *about* mental states, and these (of course) cannot be known *a priori*. *A priority*, unlike truth and falsity, cannot be reintroduced by an ascension of levels.

Consequently psychologism in general arises out of the state/content confusion. It is most obviously false in the case of logic, where structures do not quantify over objects and are not known *a posteriori*. Psychologism in logic demonstrates the falsity of psychologism in general, even though other disciplines are about things in the world and proceed *a posteriori*.

2.2. COMPOUNDING THE CONFUSION

Two factors compound the confusion between what is cognitive and what is not.

The first factor is the complex semantics of "knowledge". "Knowledge", like belief, is ambiguous between state and content. But the content-knowledge of an abstract object can actually *be the object itself*: content and object can be one and the same thing. Here are some examples:

– John knows that Mae West said at time *t* "Come up and see me some time". John's knowledge of what Mae West said at time *t* is "Come up and see me some time". So what Mae West

said and John's knowledge of it are one and the same thing – "Come up and see me some time".

– The history contained in a history book and the author's knowledge of the history contained in that book are one and the same: they are the historical propositions contained in the book.

– I know the rules of Scrabble and I write them down. I have now written down my knowledge of the rules of Scrabble – and what I have written down are the rules themselves. My knowledge of the rules is not a set of propositions *about* the rules (such as "there are 20 of them", "they are difficult"). It is the rules themselves.

Because content-knowledge of an object can *be* that object, and because state and content are commonly confused, it is easy to confuse state and object. The rules of Scrabble, for example, become confused with my knowledge of them, where my knowledge is taken to be a cognitive *state*: the rules are seen as psychological entities, as things in the head.

This confusion between state and object can be compounded by introducing an Ideal Knower, such as Chomsky's Ideal Speaker or John Macnamara's Ideal Thinker (Macnamara (1986)). Chomsky says that linguistics is concerned with the competence of an Ideal Speaker. Macnamara uses Chomsky's framework in his competence theory of human reasoning and replaces the notion of an Ideal Speaker with that of an Ideal Thinker.

Now, to say that linguistics should study the knowledge of the Ideal Speaker ("who knows its language perfectly") is to say no more and no less than "linguistics should study the language itself". This is the same kind of spurious idealisation that

we find in the claim that we should study the universe as God sees it, which is a redundant way of saying that we should study the universe as it really is. God drops out of the equation, and so does the concept of *any* Ideal Knower.

The concept of an Ideal Knower cements into place the identification of the knowledge of an object with the object itself. For instance, it identifies a language with the Ideal Speaker's knowledge of the language. Chomsky's use of "competence" sometimes means one, sometimes means the other, and sometimes means both of these things. A proper analysis splits between them and frees us from the illusion that in studying "the object of knowledge of an Ideal Knower" we are studying Mind, or in any sense doing psychology.

2.3. REVERSE PSYCHOLOGISM

Reverse psychologism is the mirror image of psychologism. Both stem from state/content or state/object confusions and both are compounded by spurious idealisation. But whereas psychologism tries to study the contents or objects of thought by studying mental states, reverse psychologism tries to study the mind by looking at the contents or objects of thought, or tries to generate cognition by internalising the public symbolisms that express the relevant content.

The best way to understand reverse psychologism is to work through some examples. I will consider two. The first is Chomsky's mentalism during his Standard Classical period, and the second is the modelling of subtraction skills in cognitive science.

3.1. CHOMSKY'S MENTALISM

Chomsky's mentalism during his Standard Classical period[5] appears to be (and is generally believed to be – see e.g. Katz (1981), Macnamara (1986)) psychologistic. In characterising mentalism he says "linguistic theory is mentalistic, since it is concerned with discovering a mental reality underlying actual behaviour" (1965, p. 4). These mental realities are the "actual subject matter of linguistics". (*ibid.*, see also pp. 193, 194) Later he says "Linguistics ... is simply the subfield of psychology that deals with these aspects of mind." ((1968), p. 24.) "I will concentrate here on some of the issues that arise when we try to develop the study of linguistic structure as a chapter of human psychology." ((1968), p. 59).

But when we look more closely we find that his approach is *reverse* psychologistic. A psychologistic position would say:

> The native speaker knows the grammar of the language. Therefore the grammar is part of her knowledge. Psychology studies human knowledge. Therefore psychology can study the internalised grammar. Therefore linguistics is a sub-field of psychology.

But Chomsky says:

> The native speaker knows the grammar of the language. Therefore the grammar is part of her knowledge. Linguistics studies grammars and can therefore tell us about the speaker's knowledge. Therefore linguistics can tell us what is going on in the head.

---

[5]Roughly from *Aspects of the Theory of Syntax* (1965) to *Language and Mind* (extended version, 1972).

The first passage goes from talking about a grammar as a content or object of knowledge to talking about a *state* of knowledge (that can be studied by psychology). The second goes from talking about a state of knowledge to an object or content of knowledge. This is reverse psychologism.

What has happened? Chomsky gives two main reasons for his mentalism.

The first is that he believes that only mentalism can account for the speaker's ability to produce and understand indefinitely many sentences. The child is exposed to a finite number of sentences, many of them degenerate, yet within a comparatively brief period acquires the ability to produce and understand indefinitely many new and well-formed ones. Chomsky says this can only be explained by saying that the child "internalises" a body of rules that gives it this ability.

The second reason he gives for his mentalism is that a speaker may not initially understand a sentence, or may not recognise an ambiguity, but may be coaxed into doing so without being given fresh information. Chomsky says that we can only explain this by saying that the speaker has an imperfect access to an *internalised grammar* that assigns these readings to the sentence. "Few hearers," he says, "may be aware of the fact that their internalised grammar in fact provides at least three structural descriptions for ["I had a book stolen"]." ((1965), pp. 21-22.)

Now, these abilities may show that the child/speaker has implicitly grasped the grammar of the language, or at least that her implicit knowledge is not accurately reflected in her performance, but it does not follow that she has *internalised* the grammar, any more than the fact that I have grasped the rules of Brazilian Canasta (or Persian Rummy, for that matter) means that I have internalised the rules of Brazilian Canasta. I

regularly forget the rules of Brazilian Canasta and have to look them up. It is not that the rules sometimes exist in my head and sometimes do not. It is that sometimes I know them and sometimes I do not. To say that the user "internalises" the grammar is to say that the grammar is a mental entity. In the same way, the rules of Brazilian Canasta would be mental entities, which clearly they are not. Neither the grammar nor the rules of the game can be destroyed by destroying everyone who knows them, yet a strong mentalist is committed to saying that they can.

Chomsky's mistake is to confuse object-knowledge with state-knowledge, or knowing. He first identifies the rules of a grammar with an (idealised) content-knowledge of them. Then he confuses this knowledge (now identified with the grammar) with knowing or state-knowledge, thus locating the grammar in the head.

The picture is compounded by his use of idealisation. He maintains that linguistics is concerned with the competence of an Ideal Speaker in a completely homogeneous speech community, and that it is only under this idealisation that performance is a direct reflection of competence and that the subject-matter of linguistics (the internalised rule-set) is available to us[6]. But we have seen that the concept of an Ideal Knower is a redundant device that falls away under analysis. To say that we should study the rules internalised by an Ideal Knower is like saying that we should study the universe as God sees it. It amounts to saying that we should study the rules themselves, the rules as they really are. When we realise this, the illusion

[6]The most commonly cited reference is (1965), p. 3, but see also (1980), (1984).

that we can do linguistics and psychology *at the same time* falls away.

I will now look at reverse psychologism in cognitive science by looking at the diagnostic modelling of subtraction skills.

2.3.2. THE DIAGNOSTIC MODELLING OF SUBTRACTION SKILLS

Diagnostic modelling is a method used in the construction of Intelligent Tutoring Systems. "Overlay" or "differential" models represent the student's knowledge as a subset of the knowledge of a hypothetical Expert, so that the student is depicted as thinking in the same way as the Expert, but as knowing less. "Diagnostic" models recognise that the student may think differently to the Expert and have misconceptions rather than a mere lack of knowledge. Diagnostic models of subtraction skills (e.g. Young & O'Shea (1982)) construct a model of what a hypothetical Expert knows and then perturb it in the hope that this will generate characteristic human errors. The model of the Expert performs atomic tasks such as "compare", "borrow", "pay back", and "add 10", and the program provides a running report of the subskills it is performing. This is seen as "looking in the mind of the Expert". Characteristic errors can be generated by perturbing the program. For example, children often fail to borrow when the top digit is lower than the bottom one. Instead they subtract the lesser number from the greater. According to the theory, they are running a procedure from which "borrow", "pay back" and "add 10" have been omitted. The Expert's program can be modified to do the same thing.

Such modelling attempts to provide cognitive models by modelling our manipulation of a public, communicable symbolism that embodies the *content* of cognition. When we watch one of these programs running we see the manipulation of numbers according to rules: the units in the unit column are compared, 10 is borrowed from the 10s column and added to the top number in the units column, and so on. This is reverse psychologism. It models the content or object of a psychological process, not the process itself.

Of course, such models might genuinely model the way in which *numbers are manipulated by us.* Some people do subtraction by decrementing the top number in the 10s column after they have borrowed. Others "pay back" by adding to the bottom number. Some people "think in blocks" (to subtract 378 from 432, subtract 378 from 400, subtract 400 from 432, and add the results). These are differences that a number-manipulating model can capture. But the models capture the content of cognitive processes, not the processes themselves.

Another way of looking at this is to say that the notion of a cognitive model is ambiguous between a *model of the student* and a *model used by the student.* A model used by the student embodies such things as her perception of the problem and her perception of how symbol-structures can be manipulated according to rules and procedures to solve the problem. The model outlined above is a model in this sense. We can say of it as it runs "this is the way in which the student believes that symbols should be manipulated to get the solution". Properly speaking it is our model of the model used by the student. It is not a *model of the student herself,* of her acts, states or processes.

The concept of the Expert plays its usual role in compounding the confusion. We are told that the Expert is an *Ideal Knower* (Miller (1982); Burton (1982)), so we would expect it

to be a device for importing "out there" structures into the head. And this is what we find. The knowledge of the Expert "merely provides a computational machine that performs the skill and is of no particular interest" (Burton & Brown (1978)). It "is not meant to be a cognitive construct, but simply a framework for relevant pieces of information" (Burton (1982)). Yet we are told that the misconceptions of the skill are represented in a network that is *psychologically real* (Burton & Brown (1978)). Young & O'Shea (1982) call the claims to psychological reality "strong claims". Thus *cognitive structures* (states, processes etc.) ostensibly emerge as perturbations of a perfect, "out there", body of rules! In fact the Expert's knowledge is just *the rules and procedures themselves.* There is nothing psychological about it. The student model is an impoverished or deviant version of these rules and procedures, and there is nothing psychological about it either.

3. BACK TO THE CHINESE ROOM

3.1. KNOWLEDGE REPRESENTATION REVISITED

Now let us return to the Physical Symbol System Hypothesis (PSSH) and the Knowledge Representation Hypothesis (KRH), and see (a) how they confuse cognition and content, and (b) how this drives the symbol-handling claim. I will look briefly at the PSSH and then look at the KRH in more detail.

Newell & Simon (1976) make it perfectly clear that the "expressions" or "symbol structures" of a physical symbol system are interpreted. They call this "designation". A physical symbol system "exists in a world of objects wider than just these symbolic expressions themselves". This notion of designation,

they say, is central to expressions, symbols and objects. They continue: "An expression designates an object if, given the expression, the system can either affect the object itself or behave in ways depending on that object. In either case, access to the object via the expression has been obtained, which is the essence of designation."

Thus, although the symbols of a physical symbol system are identified in terms of their formal or morphological properties, and manipulated according to formal rules, they are interpreted and contentful. The implications of this become clearer when we look at the KRH.

Brian Cantwell Smith says of the KRH:

> It is widely held in computational circles that any process capable of reasoning intelligently about the world must consist in part of a field of structures, of a roughly linguistic sort, which in some fashion represent whatever knowledge and beliefs the process may be said to possess. For example, according to this view, since I know that the sun sets each evening, my "mind" must contain (among other things) a language-like or symbolic structure that represents this fact, inscribed in some kind of internal code. (1985.)

Additionally, the syntax or morphology (Cantwell Smith calls it the "spelling") of this internalised symbolic structure is presumed to play a causal role in the generation of intelligent behaviour. This gives us the full statement of the KRH:

> Any mechanically embodied intelligent process will be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantical attribution, play a formal but

causal and essential role in engendering the behaviour that manifests that knowledge.

Thus a system knows that *p* if and only if it contains a symbol structure that *means* **p** *to us* and that causes the system to behave in appropriate ways. It knows, for instance, that tigers bite if and only if it contains a structure such as "Tigers bite" that causes it to climb trees in the presence of tigers. Cantwell Smith goes on to distinguish between a weak and a strong version of the KRH, and he is sceptical of both of them[7].

The story is a familiar one. The KRH does not distinguish between knowing (as a state) and knowledge (as the content or object of a state). It is assumed that to know something is to have an internalised data structure, and to have mental states is to have "a set of formal representations" (p. 37). This is the now familiar move of treating knowing as knowledge in the head.

Cantwell Smith reports that the AI community is divided in its opinion of what these internalised structures stand for. He expresses surprise at the outcome of a survey which shows that most respondents believe that the structures represent *the world*, or *situations*, rather than facts or propositions about the world or situations[8].

---

[7]Cantwell Smith does not subscribe to the KRH himself, and throughout the (1985) paper he emphasises the need to analyse and clarify concepts and issues in knowledge representation. His (1991) provides an excellent analysis.

[8]There is a striking similarity between this situation and the dilemma facing advocates of the Correspondence Theory of Truth. The latter say that the main sorts of things that are true or false are sentences or propositions, and that these are true if and only if they

This confusion arises from characterising knowledge (i.e. knowing) as internalised content. Inasmuch as internalised knowledge is seen as state or structure it can be a model and can represent a state of affairs. It has no meaning in itself, but is given one by accompanying declarative knowledge that maps it onto the state of affairs. If on the other hand it is seen as something akin to an inner sentence, then it is not a model, but it *does* express a proposition. If we run these readings together we will say that the symbol system *represents a proposition* . This is a confusion between an uninterpreted state or structure that can be used as a model to represent something, and something like a sentence, that expresses a proposition. There is, after all, no reason to believe that sentences *represent* anything. The early Wittgenstein worked such a notion hard with his picture theory of meaning, and ultimately it failed. Equally, there are no grounds for believing that *anything at all* can *represent* a proposition – though some things (such as sentences) can express them. This is another case of identifying state and content.

---

correspond to the facts. But what are 'the facts'? There are two accounts. One says that facts are 'what are expressed by sentences or propositions'. This is circular, since it is now being claimed that a sentence or proposition is true if and only if it corresponds to what it expresses. The other account avoids this circularity by saying that facts are not linguistic entities but are 'states of affairs' (sometimes called *'Sachverhalten'*, after Wittgenstein's use of the term in the *Tractatus*). This leads to a bloated ontology, for now we have to talk about not only states of affairs, but negative states of affairs, states of affairs inside other states of affairs, hypothetical states of affairs, and so on.

The KRH faces a similar dilemma: 'Do knowledge structures represent propositions or meanings, or do they represent states of affairs?' (cf. 'Do sentences correspond to propositions or meanings, or do they correspond to states of affairs?')

This becomes clearer when we look at the second clause of the KRH – that the internalised symbol structure plays a causal role in generating intelligent behaviour. The KRH tries to have it both ways: the symbol structures associated with knowledge are at once meaningful to us and causally efficacious for the system. This is the standard state/content confusion: states, but not contents, are causally efficacious; contents, but not states, are meaningful. The state of knowing that tigers bite might cause me to exhibit intelligent behaviour in the presence of tigers (such as climbing a tree), or to say intelligent things about them in their absence (such as "When in their presence, get up a tree"). But the *content* of my knowledge is not causally efficacious: the proposition expressed by the sentence "Tigers bite" cannot cause anything.

Now, this argument cheats a little, because the KRH does not say that knowledge structures cause intelligent behaviour by themselves. It says that they play a *role* in the causal process: intelligent behaviour is caused by a combination of knowledge structures *and the procedures that act upon them.*

Let us look at the history. AI has discovered that intelligence requires knowledge: for a system to be intelligent it must know a great deal. But AI does not distinguish between knowing and knowledge, and it assumes that a system knows if it contains a representation of the content of knowledge and if it has procedures that can act upon that representation, such that, together, these produce intelligent behaviour.

We have had ways of *representing what is known* for a long time. First there was speech, then there was writing, then there were libraries, now there are databases. Books, libraries and databases have knowledge represented within them. But no-one believes that they know and are intelligent. The KRH proposes something like a fancy book that responds to input on the ba-

sis of the causal efficacy of internal structures that express the content of knowledge – structures that are meaningful to us but not to the machine.

The ambiguity of "knowledge" carries over into the ambiguity of "knowledge structure", which is ambiguous between "cognitive structure" and "data structure". Knowledge structures are commonly regarded as data structures, possibly accompanied by search algorithms. Barr & Feigenbaum say, "In AI, a *representation of knowledge* is a combination of data structures and interpretive procedures..." ((1981), p. 143). Elaine Rich: "we will discuss a variety of knowledge structures. Each of them is a data structure in which knowledge about particular problem domains can be stored." ((1983), p. 203). Tore Amble: "A picture of tomorrow's computer vocabulary can be imagined, if all the words containing "data" or "information" are replaced by the word "knowledge". ((1987), p. 11) Once we have replaced "knowledge" by "data" it is easy to regard cognitive structures as data structures in the head.

There are two distinct questions. One is "How, in principle, can we construct machines that know?". This is not a technological question. It is a philosophical question. Traditional epistemology has asked "Under what conditions does agent *A* know that *p* ?" The standard answer, that philosophers have never been entirely happy with, is "*A* knows that *p* if and only if *A* believes that *p*, *p* is true, and *A* has grounds for believing that *p*": knowledge is justified true belief [9]. "How can we construct machines that know?" amounts to: "Under what conditions does machine *M* know that *p*?" This is a question in what we might call "machine epistemology".

---

[9]See Gettier (1963) for the classic list of counter examples to this claim.

Knowledge Representation asks another question: "How should we represent knowledge in machines?" This is a practical, engineering question that assumes an answer to the first question. I have argued that that answer is wrong.

3.2. CONTENT, COGNITION AND SYMBOLS

The Chinese Room Argument, then, is aimed at the symbol handling paradigm, but when we look at the hypotheses underlying that paradigm we find that they talk, not about getting meaning and understanding out of meaningless symbols, but of getting cognition by internalising and manipulating symbols that express the *content* of cognition – and that is the state/content confusion in a new incarnation.

The state/content distinction is a broader and deeper issue than that of symbol handling and cognition. For one thing, it is a special case of the general distinction between what is cognitive and what is not, and this distinction need not involve symbols at all. This is the case with the distinction between state and object, such as the distinction between my love as a state and my love as an object.

In fact we do not need symbols even in the case of content. Adrian Cussins (1990) distinguishes between what he calls "conceptual and non-conceptual content". When we say that Jo believes that Fred is a bachelor, we attribute the concept "bachelor" to Jo. But when we say that Fido thinks that the sound came from the south, we do not attribute the concept "south" to Fido. We can talk about the content of Fido's thought without attributing the concept to him, let alone the symbols that express that concept.

Now connectionism arguably models non-conceptual content. NETtalk (Sejnowski & Rosenberg (1986)), for instance, learns to pronounce words, but does not have the concept of a vowel, consonant or word. Be that as it may, connectionism maps content directly onto the nodes and activation patterns of trained-up networks.

Here is an example. There is a well-known system, due to Geoffrey Hinton, that learns family relations by back propagation (Hinton (1985), (1986)). Once trained up, the network will, for instance, give the output "Christopher" for the inputs "Penelope" and "husband". "Penelope" now has a unique activation pattern in the trained-up net, and the literature talks variously about this pattern having meaning or content, or being a representation, and so on. It is sometimes referred to as "content addressable memory". Now let *S* be the sentence "I am in Belgium" and *T* be the sentence "It is Tuesday". Let us train a net to output *T* if and only if it receives *S* as input. *S* and *T* now have unique activation-patterns, but to say that these patterns are the *meanings* or *content* of *S* and *T* is philosophically naive. It is reminiscent of Locke's claim that the meaning of a word is an idea in the head. The consequence of such a claim is that we would never understand the meaning of a word, for we have no independent access to the ideas in a speaker's head. Locke got it exactly back to front: in fact we know the idea in a speaker's head because we understand the (public) meaning of what they say, not vice versa. And it is the same with saying that meaning or content is nodes or activation patterns. But I do not need to establish this claim. My point is just that connectionism tries to internalise *content*. The major issue is the complex relationship between content and cognition, not the relationship between cognition and symbols.

BIBLIOGRAPHY

AMBLE, T. (1987). *Logic Programming and Knowledge Engineering.* (Wokingham, Addison-Wesley).

BARR, A & FEIGENBAUM, E. A. (1981). *The Handbook of Artificial Intelligence,* vol. I. (Reading Mass., Addison-Wesley).

BURTON, R. (1982). Diagnosing Bugs in a Simple Procedural Skill, in Sleeman, D. & Brown, J., (eds.) *Intelligent Tutoring Systems.* (London, Academic Press).

BURTON, R. & BROWN, J. (1978). Diagnostic Models for Procedural Bugs in Basic Mathematical Skills, *Cognitive Science,* 2.

CHARNIAK, E. & MCDERMOTT, D. (1982). *Introduction to Artificial Intelligence.* (Reading, Mass., Addison-Wesley).

CHOMSKY, N. (1965). *Aspects of the Theory of Syntax.* (Cambridge, Mass., MIT Press).

———. (1968). *Language and Mind.* (New York, Harcourt, Brace & World). Extended edition, 1972.

———. (1969). Some Empirical Assumptions in Modern Philosophy of Language, in S. Morgenbesser, P. Suppes & M. White, (eds.) *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel.* (New York).

———. (1970). Problems of Explanation in Linguistics, in R. Borger & F. Cioffi, (eds.) *Explanation in the Behavioural Sciences.* (New York, Cambridge University Press).

———. (1980). *Rules and Representations.* (Oxford, Basil Blackwell).

———. (1984). Changing Perspectives on Knowledge and Use of Language. Paper presented at a Sloan Conference, MIT, May 1984.

———. (1988). *Language and Problems of Knowledge: The Managua Lectures.* (Cambridge MA, MIT Press).

CHURCHLAND, P. M. & CHURCHLAND, P. S. (1990). Could a Machine Think? *Scientific American,* 262: 1.

CUSSINS, A. (1990). The Connectionist Construction of Concepts, in Boden, M. (ed.) (1990), *The Philosophy of Artificial Intelligence.* (Oxford, Oxford University Press).

DENNETT, D. (1987). *The Intentional Stance.* (Cambridge MA, MIT Press).

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

FREGE, G. (1967). *The Basic Laws of Arithmetic.* (Berkeley, University of California Press).

GETTIER, E. (1963). Is Justified True Belief Knowledge? *Analysis,* 23.

HINTON, G. (1995). Learning in Parallel Networks, *Byte*, April, pp.265-273.

———. (1996). Learning Distributed Representations of Concepts, *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. (Amherst MA.) pp.1-12.

HUSSERL, E. (1962). *Ideas – General Introduction to Pure Phenomenology*. (New York, Macmillan).

———. (1970). *Logical Investigations, I.* (New York, Humanities Press).

KATZ, J. (1981). *Language and Other Abstract Objects.* (New Jersey, Rowman & Littlefield).

MACNAMARA, J. (1986). *A Border Dispute: The Place of Logic in Psychology.* (Cambridge MA, MIT Bradford).

MILL, J.S. (1843). *A System of Logic.* (London).

———. (1865). *Examination of Sir William Hamilton's Philosophy.* (London).

MILLER, M. (1982). A structured Planning and Debugging Environment for Elementary Programming, in Sleeman, D. & Brown, J., eds, *Intelligent Tutoring Systems.* (London, Academic Press).

NEWELL, A. & SIMON, H. A. (1976). Computer Science as Empirical Enquiry: Symbols and Search, *Communications of the Association for Computing Machinery*, 9, March, pp.

113-126. Page references are to its reprint in Haugeland, J. (ed.), *Mind Design: Philosophy, Psychology, Artificial Intelligence.* (Cambridge MA, MIT/Bradford), 1981.

RICH, E. (1983). *Artificial Intelligence.* (Auckland, McGraw-Hill).

SCHANK, R. C. & ABELSON, R. P. (1977). *Scripts, Plans, Goals and Understanding.* (Hillsdale, Laurence Erlbaum Associates).

SEARLE, J. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences,* 3.

———. (1990). Is the Brain's Mind a Computer Program? *Scientific American,* 262: 1.

SEJNOWSKI, T. & ROSENBERG, C. (1986). NETtalk: a Parallel Network that Learns to Read Aloud, *John Hopkins Electrical Engineering and Computer Science Technical Report, JHU/EEC-86/01.*

SMITH, B. C. (1985). Prologue to Reflection and Semantics in a Procedural Language, in R. Brachman & H. Levesque, (eds.), *Readings in Knowledge Representation.* (Los Altos, Morgan Kaufmann).

———. (1991). The Owl and the Electric Encyclopedia, *Artificial Intelligence,* 47.

WITTGENSTEIN, L. (1961). *Tractatus Logico-Philosophicus.* (London, Routledge & Kegan Paul). (Original German edition published 1921.)

WENGER, E. (1987). *Artificial Intelligence and Tutoring Systems.* (California, Morgan Kaufmann Publishers).

YOUNG, R. & O'SHEA, T. (1982). Errors in Children's Subtraction, *Cognitive Science*, 5.