J. Expt. Theor. Artif. Intell. 1(1991)283–296.

*Research Paper*

# Searle's AI program

KEVIN B. KORB

*Department of History and Philosophy of Science, Indiana University, Bloomington, Indiana 47405*
email: kkorb@ucs.indiana.edu

*Abstract.* John Searle has used his Chinese room example to attack the idea of computationally reproducing intelligence. His arguments have variously assumed or (more recently) asserted that consciousness and intelligence are necessarily interdependent. This stance has allowed him to apply intuitive arguments about what could or could not be conscious to the issue of what could or could not be intelligent. I present a variety of arguments, theoretical and intuitive, to show that Searle is conflating mentality and semantics. By maintaining that distinction we can then address how to generate the semantics that intelligence requires. In Stevan Harnad's approach to symbol-grounding we have a plausible candidate for finding referential semantics without taking detours through an unanalysable consciouness. Artificial intelligence as normally construed does not require that philosophical problems about consciousness be resolved, let alone that consciousness should be computationally definable: Searle's arguments against strong AI are irrelevant to real-world AI.

*Received 30 November 1990; revised 9 July 1991*

## 1. Overview

Artificial intelligence (AI) has struggled since its inception with the problem of justifying its existence. John Searle's arguments against the possibility of success for a strong program in AI have put the problem in an especially acute form and so have preoccupied many of the foundational discussions. Searle claims that AI suffers from confusion at its core: that the very idea of computationally reproducing intentionality is incoherent.

I shall argue in part that Searle's perspective is distorted by a failure to distinguish two clusters of concepts: one centred on consciousness (e.g. mentality, raw feels, phenomenology of cognition) and the other on intentionality and semantics. Searle himself says (1990, p. 30):

> [My] argument rests on the distinction between the formal symbol manipulation that is done by the computer and the mental contents biologically produced by the brain, a distinction I have abbreviated—I hope not misleadingly—as the distinction between syntax and semantics.

You will notice that Searle has equated syntax with symbol manipulation on the one hand and semantics with mentality on the other. It is this latter identification which I suggest has misled Searle.

According to Searle, AI comes in two flavours: strong and weak. Both flavours propose to achieve an understanding of intelligence through simulating it on a digital computer. Strong AI claims that we can produce that understanding within

*the machine*: it claims that simulating the mind is tantamount to recreating it. Weak AI, by contrast, claims that this understanding would be produced (only) within *our contemplation* of that simulation; the simulation itself may be as detailed as you like, but it could no more *be* intelligence than astrophysical simulations could be a sun.

It is unsurprising that the AI community should reject this weak AI, for then AI would be a mere computational research tool within cognitive psychology; it would have no more claim to being an independent discipline than do computational approaches to astrophysics.[1] What is surprising, however, is the failure to point out that there are other conceivable approaches to AI beyond, or between, these two. In particular, I argue that there is a coherent *moderate* AI holding that *intelligence* can be produced computationally without necessarily (by the same process) producing mentality or consciousness. By refusing to identify mentality with cognition or semantics, we can adopt the more modest goal of recreating intelligence without addressing the problem of consciousness. And by refusing to speak to the problem of consciousness, Searle's strong intuitions about where the seat of consciousness might be simply are no longer pertinent.

Figure 1 lays out some of the available views on digital simulation and its capability for inducing mentality and intentionality. Here I define 'strong simulation' as the view that simulation reproduces the process concerned and 'weak simulation' as the opposite view. 'Moderate simulation' expresses the idea that digital simulation plus *something else* suffices to reproduce the process. The vertical dimension shows the distinction between intentionality and mentality.

It turns out that my intermediate AI is not quite intermediate between strong and weak AI; that position seems to be occupied by Stevan Harnad. Harnad apparently believes that semantics and consciousness are inseparable, but also that neither can be attained without symbol grounding (see Harnad 1989*a*). Since I buy into symbol grounding (more of which later), my position is indeed moderate, but, as you can see from the chart, it is agnostic (so far) regarding the mentality of AI systems.

Figure 2 outlines the main points behind my discussion. Searle's core argument

| | Simulation | | |
|---|---|---|---|
| | Weak: Computation merely simulates | Moderate: Computation + X recreates | Strong: Computation recreates |
| Semantics, Intentionality, Cognition | ? | Korb | ? |
| Mentality + above | Searle | Harnad | Strong AI |

Figure 1. What are the proper aspirations of AI?

is (1), that intelligenc
consciousness. He mad
Argument, where synta
argument simply *assum*
Searle took as conclusiv
*feel* any understanding,
his latest contribution to
fill in this second link,
'Syntacticism' refers t
Room Argument and i
however, is the accepta
showing that semantics
My view (3) is that S
syntax does not suffice f
intuitive arguments can
does require semantics,
consciousness (1b). By
from consciousness as i
outlook for simulating c
on AI itself.

**2. Semantics without a**
Although semantics, re
interrelated) concepts, t
any nice discriminations
as my test case. My relia
semantics will suffice fo
this distinction plausible

(1) Searle: Intelligence –
(1a) Intelligence –
*Th
(*a*
(1b) Semantics →
*Th
(2) 'Syntacticism' accept
out of syntax, b
(2a) The Turing
or (2b) Truncated (r
or (2c) Enough synta
(3) Syntax does not suffi
(3a) Intelligence –
(3b) Not: Semanti
● Harnad's
mod
● Searle's a
gooc

tantamount to recreating it.
g would be produced (only)
simulation itself may be as
elligence than astrophysical

eject this weak AI, for then
thin cognitive psychology; it
cipline than do computational
vever, is the failure to point
I beyond, or between, these
t *moderate* AI holding that
ut necessarily (by the same
efusing to identify mentality
e modest goal of recreating
iousness. And by refusing to
g intuitions about where the
pertinent.
n digital simulation and its
ity. Here I define 'strong
the process concerned and
imulation' expresses the idea
reproduce the process. The
entionality and mentality.
intermediate between strong
by Stevan Harnad. Harnad
ss are inseparable, but also
ng (see Harnad 1989*a*). Since
ter), my position is indeed
gnostic (so far) regarding the

ssion. Searle's core argument

ion

| te: tion es | Strong: Computation recreates |
|---|---|
| | ? |
| | Strong AI |

rations of AI?

is (1), that intelligence requires semantics and semantics in turn requires consciousness. He made the first step in the chain with the Chinese Room Argument, where syntax turned out to be insufficient for understanding. This argument simply *assumed* that semantics and consciousness are interdependent: Searle took as conclusive evidence that there was no semantics that he could not *feel* any understanding, that he did not have the right state of consciousness. With his latest contribution to *Behavioral and Brain Sciences* (forthcoming) he has tried fill in this second link, as we shall see.

'Syntacticism' refers to a miscellany of views in opposition to Searle's Chinese Room Argument and in defence of strong AI. What they share with Searle, however, is the acceptance of (1). Problems are avoided by somehow or other showing that semantics is reducible to syntax.

My view (3) is that Searle is right when he says that it is a logical truth that syntax does not suffice for semantics. Rather than argue to the contrary, Searle's intuitive arguments can be defused in a more plausible way: whereas intelligence does require semantics, there is no reason to believe that semantics requires consciousness (1b). By breaking the latter link, I can dismiss Searle's arguments from consciousness as irrelevant; and I could even adopt a 'weak' view of the outlook for simulating consciousness without in any way compromising my views on AI itself.

## 2. Semantics without awareness

Although semantics, reference, intentionality, intelligence, etc. are distinct (if interrelated) concepts, the bulk of the discussion has proceeded without making any nice discriminations here. And I shall do the same, using referential semantics as my test case. My reliance on the grosser distinction between consciousness and semantics will suffice for some modest progress, so now I will attempt to make this distinction plausible.

(1) Searle: Intelligence → Semantics → Consciousness
  (1a) Intelligence → Semantics
    *The Chinese Room Argument
    (*assumed* semantics = consciousness)
  (1b) Semantics → Consciousness
    *The Sleeping Man Argument
(2) 'Syntacticism' accepts (1), but rejects the claim that AI can't get semantics out of syntax, because either:
  (2a) The Turing Test *defines* intelligence
  or (2b) Truncated (non-referential) semantics is good enough (Rapaport)
  or (2c) Enough syntax *just is* semantics (Pylyshyn, the Churchlands)
(3) Syntax does not suffice for semantics. But not to worry, because, although
  (3a) Intelligence → Semantics
  (3b) Not: Semantics → Consciousness
    - Harnad's symbol grounding provides reason for optimism regarding moderate AI
    - Searle's arguments regarding consciousness may or may not be good, but are in any case irrelevant to moderate AI

Figure 2.

The game begins, as always, with Searle's 'decisive refutation' of strong AI, namely his renowned Chinese Room Argument (Searle 1984, pp. 31–33). In the Chinese room is an English speaker, say Searle himself, who understands no Chinese. The room has an input/output slot. When a Chinese symbolic expression comes into the room, Searle is to look it up in a rule book (written in English) which tells him what Chinese symbols to output (if any). The rules are based strictly on the syntactic features of the Chinese symbols; no semantics are revealed. If this room were put to the Turing Test, it would pass—that is, it is assumed that the set of rules are sufficient for that to happen (if we can ignore the excessive time required to generate a response). But there is no understanding of Chinese going on here!

One point of this thought-experiment is that the Turing Test is insufficient to establish the presence of intelligence. But the implications are intended to go far beyond that. What Searle takes his argument to reveal is indicated by his claim that all possible counter-arguments are *necessarily* inadequate 'since the argument rests on a very simple logical truth, namely, syntax alone is not sufficient for semantics, and digital computers insofar as they are computers have, by definition, a syntax alone' (1984, p. 34). There is at least one variety of counter-argument which Searle overlooks—one that points out that he has the definition of 'computer' wrong.

Early in *Minds, Brains and Science* (1984), Searle lists four features of mental phenomena which are so basic that they can act as a litmus test for any philosophy of mind; i.e., a theory which does not allow for them is *ipso facto* inadequate:

> *Consciousness*. We have conscious mental states.
> *Intentionality*. We have intentional states, that is states which refer to objects or states of affairs other than themselves.
> *The subjectivity of mental states*. Our mental states are not *equally* accessible to external observers.
> *Mental causation*. Our thoughts and feelings have some causal effect on the physical world (e.g. our behaviour).

I might have some qualifications to make, but basically I agree that these are reasonable requirements to impose on a general theory of human mentality.[2] Less obvious, however, would be the claim that these *same* requirements describe the narrower domain of intelligence *per se*. I'm proposing here that intentional states—states that have semantic content—are not necessarily mental states. Opposing this view, Searle asserts: 'It is easy enough to imagine a universe without [consciousness], but if you do, you will see that you have imagined a universe that is truly meaningless' (1984, pp. 15–16). Perhaps in some metaphorical and soul-stirring sense of 'meaningless' Searle is right; but if he literally means the world must contain no semantics, then he literally is wrong. *This* world (that is, this planet) can be readily made-over into a world without consciousness—as we know full well—while leaving intact a large number of semantic-bearing artifacts. Indeed, quite a few early historical recordings clearly have semantic content, although there is no consciousness anywhere which shares that content.

No doubt Searle would argue that my point is just trivial. Everyone grants that, say, the books we write (and this paper as well, however wrong-headed) have semantic content. But we do not conclude that our books *think*! What is at issue is not the existence of semantic content, but the existence of intentional

*processes*, or 'semantic
is after is exactly reproc
question, we should nc
issue *were* just the exis
anything holding symbo
when he asserts that co

**3. Are sleepy heads en**

OK, so *we* can interpre
that their 'knowledge re
*they* be intelligent if the
are not self-interpreting
about how semantics ar

Searle tells such a sto
look at his own semantic
by a phenomenological
feeling of understanding
story then is that semar
a computational accoun
computational sequel—t

Searle's argument begi
shape'; e.g. perceptions
about an object always
presents the Sleeping M

> Imagine that a man is ir
> true to say of him that
> believes that Denver is
> States, etc. But now, *wh*
> *beliefs*? Well, the only
> neurophysiological facts.
> and processes. (forthcom

But now Searle claims w
intentional] states have
facts, because there is n
The only way out, al
consciousness.

But this argument is a
the same time the only
beliefs are themselves n
they are such biological f
are neurophysiological fa
level. But then what co
level);[3] therefore, what
biological. However, Se
shapes hidden among the
facts. Since, following S
inescapable that he has

This is less an apparer
*reductio ad absurdum*! It

*processes*, or 'semantic engines.' This is, I think, the right response, for what AI is after is exactly reproducing intentional processes. But before moving on to this question, we should note that my hypothetical Searle has given ground. If the issue *were* just the existence or not of semantics, then we can see already that anything holding symbols of our invention can have a semantics. Searle is in error when he asserts that computers *by definition alone* are restricted to syntax.

### 3. Are sleepy heads empty?

OK, so *we* can interpret the symbols that we pack into computers; *we* can say that their 'knowledge representations' indeed represent something. But how can *they* be intelligent if they cannot interpret those symbols for themselves—if they are not self-interpreting systems? This I take to be a legitimate request for a story about how semantics arises.

Searle tells such a story in his new article (forthcoming). In effect, he takes a look at his own semantic phenomena and notices that they are always accompanied by a phenomenological aspect (thus, when we understand something, we have a feeling of understanding—what Tim van Gelder calls 'feelie semantics'). Searle's story then is that semantics arises only out of consciousness. Unless we can get a computational account of this primordial consciousness, this story will have no computational sequel—there will be nothing left of AI but weak simulation.

Searle's argument begins with the point that every intentional state has 'aspectual shape'; e.g. perceptions always reflect the perspective of the agent, and beliefs about an object always concern some, but not all, of its characteristics. Then he presents the Sleeping Man Argument:

> Imagine that a man is in a sound dreamless sleep. Now, while he is in such a state it is true to say of him that he has a number of unconscious mental states. For example, he believes that Denver is the capital of Colorado, Washington is the capital of the United States, etc. But now, *what fact about him makes it the case that he has these unconscious beliefs*? Well, the only facts that could exist while he is completely unconscious are neurophysiological facts... There is simply nothing there except neurophysiological states and processes. (forthcoming, section 2, step 4)

But now Searle claims we have an apparent contradiction since 'the [unconscious intentional] states have an aspectual shape that cannot be constituted by [the] facts, because there is no aspectual shape at the level of neurons and synapses.' The only way out, allegedly, is to give the unconscious beliefs *potential* consciousness.

But this argument is astounding! The sleeping man has a set of beliefs, and at the same time the only facts true of him are neurophysiological. So, either his beliefs are themselves neurophysiological facts or they are non-factual. Perhaps they are such biological facts; indeed Searle (in footnote 4) says that beliefs simply are neurophysiological facts 'at a higher level,' that they supervene on the physical level. But then what constitutes these beliefs are biological facts (at whatever level);[3] therefore, what constitutes aspectual shape in these beliefs is likewise biological. However, Searle rejects this because he cannot locate the aspectual shapes hidden among the neurons and synapses. So the beliefs cannot be biological facts. Since, following Searle, that's all there is for the unconscious man, it is inescapable that he has no beliefs at all!

This is less an apparent contradiction than a real one, a kind of self-inflicted *reductio ad absurdum*! It is a curious notion that invoking potential consciousness

will allow Searle to escape its force. These potentialities themselves are either true of the sleeping man or they are not. By Searle's arguments they will therefore turn out to be either biological facts or fantasy. Searle goes on to opt for an underlying biological reality, for he says that the beliefs are supported by objective causal features (section 2, step 6)—and if there is some non-biological causality here, Searle has failed to name it. So, the aspectual shapes of belief, hidden or not, must be embedded in the biological, causal structure. Searle's refusal to acknowledge their presence seems to be a posture designed to leave a role for consciousness.

The *reductio* need not concern *us* unduly, since it is directly attributable to Searle's abandonment of his own sensible suggestion of viewing intentional states as higher-order biological facts. And so it lacks all force to demonstrate what he wanted, that intentional states necessitate conscious states.

## 4. Brain soup

Since I am pushing the idea that intentionality and consciousness are potentially independent, I need a more positive argument beyond discomfiting Searle's. I'd like to start with some suggestive, almost anecdotal evidence. Instead of sleeping, consider dreaming. Dreams clearly (sometimes) have content; they are about something. But we are obviously not conscious when we are dreaming. One might choose to argue that there is *some* degree of consciousness involved. But even if that is true, my suggestive point remains: semantic processing is available under diminished circumstances, and there is no obvious reason why the diminishment cannot be total.

Another sort of example is offered by guidebooks on 'how to solve problems.' They often advise that when you get stuck on a problem you should set it aside, stop 'thinking about it.' Frequently, the answer will just 'pop into your head' sometime later. Now, could it really be the case that there was no thinking going on? If so, it would seem that the solution, the right thought, must have been just sitting in your head, waiting for recognition; the mental block was not in thinking through the problem, but simply in realizing that you were done! The more plausible description is that there has been thinking going on without consciousness or 'raw feel.' Again, this argument is intended merely to be suggestive—although, if only on introspective grounds, I do believe in the existence of complex, unconscious thought processes.[4]

It may be underappreciated that there is a substantial, and growing, body of experimental work showing that semantic processing goes on in humans in many cases without any conscious participation. Thus, Anthony Marcel has demonstrated the existence of a phenomenon that has been called 'unconscious reading': when words (or blanks) are presented to subjects with progressively less time exposure (before masking), at some point the subjects' ability to discern the presence or absence of words drops to the level of random guessing. But *below* that time interval the exposed word continues to cause *semantic* associative priming (see Marcel 1983). M. H. Marx, in a survey article, concludes that 'many diverse lines of inquiry converge to suggest that unconscious processing of information is a real and very general phenomenon...' (Marx 1984, p. 217).

The interpretation of such cases as showing a significant distinction between consciousness and cognition might be dismissed on the ground that they indicate only that consciousness is dispensible in some cases, that given a surrounding context of conscious processing some peripheral or vestigial nonconscious cognition

might be found as wel
interpreting mechanisms
or mentality. Recently,
how a solution might b
(1989*a*, 1989*b*). Harnad
for how some primitive
is that upon the presen
one produced by a hors
object; this projection i
superimposed, differenti
to different features of
sorted into those for w
features yield categori
specific icons as memb
categorical representati
those objects and are
image 'means' its object
in place of the categoric
is directly, causally grou
as yet unspecified, ca
language. For more det

This I take to be a ser
Nothing in it requires or
system. And that is as it
in fact does not help in
is based on an inappro
conscious of our own t
these have semantic co
external world, it is clea

But that is *our* conditi
ch. 1) shows that it is c
advantage of being able
simulating machinery w
of doppelgänger brains
identical to our own); s
at this that there is not th
and our own. These bra
world. They have no re
they seem to perceive (w
populates their world ma
function, they function e
they are conscious;[5] ind
identical in every detail
supposition of such var
anywhere in explaining t

## 5. Can syntax beget ser

To the challenge of find
mind, there is a possibl
grounding, namely that

might be found as well. So we must move to a different argument: that self-interpreting mechanisms can be described that presuppose nothing of consciousness or mentality. Recently, Stevan Harnad has sketched out how this might be done, how a solution might be found to what he calls the symbol-grounding problem (1989*a*, 1989*b*). Harnad outlines a theory which, at least potentially, can account for how some primitive category symbols can acquire meaning. Briefly, his story is that upon the presentation of an external stimulus of some natural kind, say one produced by a horse, our sensory surfaces contain a projection of the distal object; this projection is transformed into an iconic representation; icons can be superimposed, differentiated, etc. by low-level cognitive functions that are sensitive to different features of icons; sensitivity to iconic features allows icons to be sorted into those for which certain features are invariant; clusters of invariant features yield categorical representations—they are sufficient for identifying specific icons as members of a category. At this stage we have iconic and categorical representations of physical objects; these are related causally with those objects and are not yet conveyors of meaning (no more than a camera image 'means' its object). However, by using an arbitrary *symbolic* token ('horse') in place of the categorical representation we have a symbol, the meaning of which is directly, causally grounded in the world. ('Using in place of' designates some, as yet unspecified, causal relation bridging categorical representations and language. For more detail on symbol grounding, see Harnad 1989*a* and 1987).

This I take to be a serious start on giving an account of the origin of semantics. Nothing in it requires or assumes any consciousness within or without the semantic system. And that is as it should be. The assumption of a pre-existing consciousness in fact does not help in explaining the origin of semantics. The idea that it does is based on an inappropriate generalization from our own condition: we are conscious of our own thoughts and their semantic contents. How do we know these have semantic content? Well, so long as one is not skeptical about the external world, it is clear that many of our beliefs are about that external world.

But that is *our* condition. Hilary Putnam's treatment of 'brains in a vat' (1981, ch. 1) shows that it is coherent to talk of consciousnesses which do not have the advantage of being able to refer. Imagine an evil scientist equipped with world-simulating machinery which he has attached to the afferent and efferent nerves of doppelgänger brains lying in a vat of nutrients (i.e. the brains are structurally identical to our own); suppose further that the scientist-cum-machine is so good at this that there is not the slightest difference between their supposed perceptions and our own. These brains cannot successfully refer to physical objects in their world. They have no relevant causal connection with such things; and so what they seem to perceive (which are just the objects familiar to *us*) and what in fact populates their world may diverge as much you please. But these brains not only function, they function exactly as do our own. It would seem absurd to deny that they are conscious;[5] indeed, we can suppose that they have phenomenologies identical in every detail to our own. But then it follows immediately that the supposition of such varieties of consciousness as we possess does not get us anywhere in explaining the origin of referential semantics.

## 5. Can syntax beget semantics?

To the challenge of finding a starting point for semantics that does not invoke mind, there is a possible response quite different from that offered by symbol grounding, namely that syntax *alone* suffices for semantics, that is, syntacticism.

This view tends to be encouraged by the idea that the Turing Test can *replace* (or define) the concept of intelligence. For if intelligence can be reduced to an identifiable class of linguistic behaviours, then there is good reason to believe (*pace* Hubert Dreyfus) that a computer system without symbol grounding could produce them.

This argument fails because the Turing Test, however practically useful as a guide to research, is fundamentally flawed as a definition of or replacement for the concept of intelligence (Hofstadter and Dennett are recent proponents of the Turing redefinition of intelligence, 1981, pp. 94–95).[6] The most direct demonstration of this I know proceeds from a consideration of 'chess intelligence.' The Turing Test for this domain is simply for a strong human player to play the machine and lose. Of course, there already are machines that qualify, but I'm not concerned with whether or not we would grant them 'chess intelligence.' Instead, consider my own machine Randy which plays a game against Gary Kasparov (the current world champion) and not only wins it but in doing so outplays Kasparov at every stage, winning with a brilliant positional sacrifice in the endgame. Let us say then that Hans Berliner and company at Carnegie Mellon University issue a public statement announcing the supreme chess intelligence of my machine and rush down to Indiana to find out how I did it. My story, however, is embarrassingly simple: after generating all the legal moves in a position, my program randomly chooses one of them to play. Given about 35 legal moves per position, then it has a probability of $(1/35)^{50}$ of choosing the very best moves in a fifty move game. Regardless of how many times you put it to the Turing Test for chess intelligence, there is a *non-vanishing* probability of its passing (by the time it vanishes we shall all have joined Keynes).

The analogy between this and the original Turing Test is clear (something like it appears in 'folk culture' in the form of monkeys typing Shakespeare). By randomly selecting words, a Randy *could* pass the Turing Test. Nobody (I hope) would insist that the randomizing machine by virtue of its passing the Turing Test would be in fact intelligent. The rational conclusion would be instead that we live in a strangely improbable, possible world. And that conclusion alone shows the logical insufficiency of the Turing Test.

If semantics were somehow a 'higher level' of syntax, then the Chinese Room Argument would not make any sense at all. For Searle's room specifically allowed any amount or kind of syntax one cares to imagine, and therefore also any semantics which can be defined solely in terms of syntax. But the Chinese room is not so easily dismissed. Harnad (1989*a*) gives a compelling reinterpretation of the Chinese room: suppose you have a Chinese-Chinese dictionary, of whatever complexity, that makes no contact with symbols or objects without using the Chinese language (for example, there are no pictures). Could you learn Chinese from such a dictionary? It might be conceivable, if you apply pre-existing knowledge of Chinese culture, linguistics, etc. But surely it is not conceivable if this is the *first* language you must learn—you are just stuck on Harnad's Chinese-Chinese Merry-Go-Round.

William Rapaport appears to be *content* with systems that have no referential semantics, because he is concerned with *internal* semantics—'semantics as an interconnected network of internal symbols' (1988, p. 94). Symbols at least initially acquire meaning from some contact with the world, as Rapaport admits (p. 95). But thereafter they can live independent lives. The links that terminal nodes have

with the external world
some 'dialect' of English
with our AI program u
On the contrary, I th
is no such internal sema
internal semantic associ
histories and our cont
correct, add to and ref
experimentation and sc
world could *at best* be
its subsequent evolution
An additional problen
arguments against the
early) Thomas Kuhn (s
theoretical term is fully
network), then two the
do not. Einstein's rela
incorporate different r
supplying different mea
Far from being in oppos
things! (Did poor Einst
defeat an old theory by
term is *not* a full determ
is an obvious candidate
another.[7]
Since our concern her
fully comparable to hum
or to evaluate theories
the 'internal semantics'
A final argument for
viable interpretation of
the interpretation gives
because it does not even
(Pylyshyn 1984, pp. 43–
theory of reference (or s
relations do not provide
interpretation. This amo
acquire semantics if you
with the view is the idea
it forms the first word
without meaning; but if
word over the turf then
content. Not only is this
just never can be enou
singleton set. As Putnar
complicated set of senter
the while preserving the
possible worlds.[9] Not o
interpretation, hooking

with the external world can be changed in arbitrary fashion, thereby just obtaining some 'dialect' of English. It is a trivial question whether we choose to communicate with our AI program using English or using some strange dialect.

On the contrary, I think that the triviality here is merely apparent—that there is no such internal semantics *independent* of links to the external world. Our own internal semantic associations are heavily dependent upon both our perceptual histories and our continued intervention in the environment. We constantly correct, add to and refine our semantic interconnections as a result of worldly experimentation and social interaction. A semantic world without an external world could *at best* be a snapshot of our own interior semantic landscape—and its subsequent evolution could not track our own without perpetual editing.

An additional problem with this view is that it abandons one of the more telling arguments against the incommensurability thesis of Paul Feyerabend and (the early) Thomas Kuhn (see Feyerabend 1975, chapter 17). If the meaning of a theoretical term is fully determined by the role it plays in the theory (or semantic network), then two theories which ostensibly concern the same subject matter, do not. Einstein's relativity theory and Newtonian mechanics, for example, incorporate different rules at their very foundations and therefore must be supplying different meanings to their theoretical terms, such as mass, force, etc. Far from being in opposition to each other, they do not even talk about the same things! (Did poor Einstein slip up?) Either the conclusion is that we can *never* defeat an old theory by introducing a better one or that the functional role of a term is *not* a full determiner of its meaning. The extension, or reference, of terms is an obvious candidate for supporting the measurement of theories against one another.[7]

Since our concern here is the production of artificial systems with intelligences fully comparable to human intelligence, the inability to have a worldly semantics or to evaluate theories against one another is a sufficient objection to adopting the 'internal semantics' approach.[8]

A final argument for the sufficiency of syntax asserts that if there is only one viable interpretation of a batch of syntax, then that batch just *has* that meaning the interpretation gives it. Putnam calls this the 'magical theory of reference,' because it does not even put the two relata—word and object—into any relation (Pylyshyn 1984, pp. 43–44, tends towards such an approach). Against a causal theory of reference (or symbol grounding), the counter would be that the causal relations do not provide anything more than additional constraints on any possible interpretation. This amounts to the claim that a system with syntax alone can acquire semantics if you just throw more and more syntax into the pot. Consistent with the view is the idea that if a forest drops pollen onto the ground such that it forms the first word of the Encyclopaedia Britannica, then this is an event without meaning; but if the pollen spells out the entire encyclopaedia word for word over the turf then both the forest and the encyclopaedia have the same content. Not only is this notion (by my intuition) absurd, but it seems that there just never can be enough syntax to restrict all possible interpretations to a singleton set. As Putnam showed in *Reason, Truth and History*, an arbitrarily complicated set of sentences can be reinterpreted in entirely unintuitive ways, all the while preserving the truth and falsity of the sentences in the set across all possible worlds.[9] Not only is the Britannica insufficient to specify a single interpretation, hooking the Britannica up to a perceptual system is likewise

insufficient, if the 'cash value' of the perceptual system is limited to accumulating uninterpreted sentences—no matter how large the Britannica system may grow.

The intuition that a syntax complicated enough to yield a a single possible interpretation *ipso facto* has a semantics is *at best* the consequence of a limited imagination. On the other hand, adding a causal connection to the world is different in kind from throwing in another chunk of syntax. By relating symbol and object, it is at least a candidate for going beyond constraining possible interpretations to actually *introducing* some interpretation.[10]

## 6. The Roboman reply

Let us now reconsider Searle's Chinese Room Argument, while holding onto the separation of intentionality and phenomenology. Two of the common counters to the Chinese room which Searle explicitly considers are the systems reply and the robot reply (1981, pp. 288–294). The systems reply ascribes intentionality to the Chinese room as a whole: it is the conjunction of the person operating out of the rule book, the rule book itself, the paper with Chinese symbols, etc. which together understands Chinese. Searle answers that we can imagine the man internalizing the whole rule book by memorization and using memorized sounds and vocalizations instead of written symbols; but, since he is still using the rules in a rote fashion, it is a 'plain fact' that there is still no understanding. Since Searle offers no further argument, he must be relying again on the presumptive absence of the *feeling* of understanding to carry the day. As I have argued, however, feelings one way or the other are not going to decide the issue.[11] Recently this has turned into a bizarre debate about whether or not the Searlean system is manifesting a multiple personality syndrome, one Chinese-speaking and the other English-speaking. Harnad takes the position that it fails to show the classic signs of such a syndrome and so is not such a manifestation. This is open to doubt, though, since here the etiology would have to be very far from classic!

Nevertheless, we cannot conclude that the Chinese system has intentionality. The 'Chinese personality' posited by Searle's opponents suffers from some remarkable shortcomings. You will recall that the memorized Chinese room has no provision for connecting Searle's perceptions and observations to Chinese symbols. Asked about the colour of my shirt, for example, it will be unable to provide the answer. Since it is capable of passing the Turing Test (and so has some skill in dealing with tough questions), perhaps it would beg-off by suddenly announcing that it has a splitting headache and indicating a desire to leave—but, of course, it would have no means of communicating that desire to Searle, let alone compelling Searle to act on it.[12] The point is that this so-called personality does not have any contact with the external world—there is no referential semantics going on here!

The robot reply suggests that if we were to put the Chinese room inside a robot that had cameras for input and mechanical effectors for output, then these would supply the missing contact with the world and so there would be understanding. Searle responds by first pointing out that this is a retreat from the original strong AI view; it is no longer just the formal structure of the computer program which is claimed to suffice for intentionality, rather it must be supplemented by a causal connection between the symbols and the things signified. But secondly, this revision does not save the case: since the man inside the room inside the robot continues to see only the Chinese symbols come in and out, the

game he is playing cor
game. He has no new u
the scraps of paper, th
distal object is irrelevar
the symbol manipulatio

The systems and robo
room or anything in it
with the combination of
reply (1981, pp. 295–29
as any proponent of A
behaved in a seemingly i
and, in particular, woul
addition to sensory appa
uninterpreted formal sy

But here the argume
How are we supposed t
are uninterpreted or no
would have to agree tha
simple robot reply. But
the causal processes in
Chinese symbols into th
no access to the causal p
But the proper model
activities to be taken int
categorization of an imag
that *is* interpretation.

Consider again the cas
more than that (since th
causal apparatus—that is
feed his brain Chinese s
built which in turn take
that the causal relations b
irrelevant to the *system*; t
indeed those causal relat
be appropriate (for ther
Roboman will have to
characteristics in much
engineering problem onl
of Roboman just *is* Searl
homo-normal. There is n
system must be uninterpr
apparatus in the system a
and the symbols, we hav

## 7. Conclusion

Searle's earlier point, na
away from strong AI? Th
out that the AI Searle w
is totally unable to make

game he is playing continues to be nothing more than a formal manipulation game. He has no new understanding of Chinese; nor does the conjunction of him, the scraps of paper, the robot's skin, etc. The causal connection of symbol to distal object is irrelevant both to the man in the box and, more importantly, to the symbol manipulation process.

The systems and robot replies fail to give us a reason to believe that the Chinese room or anything in it understands Chinese. Where Searle is less successful is with the combination of these two objections. He does consider this combination reply (1981, pp. 295–297): he says that it would be his inclination, just as much as any proponent of AI, to ascribe intentionality to a robot-as-a-whole if it behaved in a seemingly intelligent fashion. But this ascription would be provisional and, in particular, would be revoked if he were to discover that the robot, in addition to sensory apparatus and effectors, only had a mechanism for manipulating uninterpreted formal symbols.

But here the argument bogs down; indeed, it borders on question-begging. How are we supposed to know whether the symbols that the robot manipulates are uninterpreted or not? To be sure, Searle has described one case where we would have to agree that the symbol system is uninterpreted, namely that of the simple robot reply. But that argument does not carry forward: it assumed that the causal processes introduced with the robot contribute only by dropping Chinese symbols into the room. The symbol processing subsystem (Searle) had no access to the causal processes beyond observing their final symbolic deposits. But the proper model for the combination reply allows for all the systemic activities to be taken into account, from the activation of sensory surfaces to the categorization of an image to the selection of a symbol. That is symbol grounding—that *is* interpretation.

Consider again the case where the man internalizes the Chinese rule book; but more than that (since this *is* the combination case), he internalizes the Robot's causal apparatus—that is, a camera and microphone are built into his head which feed his brain Chinese symbols, and mechanical limbs and vocal apparatus are built which in turn take Chinese symbols as their inputs. Searle cannot now argue that the causal relations between the symbols and the external world are necessarily irrelevant to the *system*; this Roboman *can* use them in building an interpretation—indeed those causal relations are *constitutive* of Roboman. For the semantics to be appropriate (for them to match that of native Chinese speakers) perhaps Roboman will have to have a perceptual psychology that selects out salient characteristics in much the same way that humans do, etc. But this is an engineering problem only so far as the camera, etc. are concerned, and the rest of Roboman just *is* Searle's neural apparatus, which cannot be too deviant from homo-normal. There is no possibility here of a general argument that the symbol system must be uninterpreted. In other words, by embedding a causal, perceptual apparatus in the system and allowing relevant connections between that apparatus and the symbols, we have potentially resolved the symbol-grounding problem.[13]

## 7. Conclusion

Searle's earlier point, naturally, remains: haven't we wandered a good distance away from strong AI? That we have—as Searle has characterized it. But it turns out that the AI Searle would have us consider is one so purely symbolic that it is totally unable to make referential contact with the world. Furthermore, the

goal of that AI is the recreation of the entire panoply of human mental life, rather than the creation of (a possibly alien form of) cognition. The vast majority of research in artificial intelligence does not presuppose anything stronger than the goal of producing *intelligence alone* and so need not be concerned with the many arguments defeating strong AI.[14]

## Notes

1. There must be some relation between cognitive psychology and AI, but that is not my topic here.
2. Searle makes heavy use of common-sense intuitions about mentality, such as these. It has perhaps become fashionable to regard such intuitions as expendable—and, so, to criticize Searle for using them. The reason given for their expendability is that common intuitions have very often in the past been overthrown by subsequent scientific advances which reveal them to have been largely founded on prejudice. Thus, the Churchlands' (1990) try to draw an analogy between Searle's intuitions and a hypothetical nineteenth-century intuition that light and magnetism have nothing to do with each other. While I concede that intuition should give way to science, I disagree with the sweeping dismissal of intuitive arguments. Without intuitions one's reasoning cannot even get started. Furthermore, there is a vast difference between discrediting an intuitive judgment through scientific research—which has some hope of being objectively constrained—and throwing out common intuition (or common sense) because it conflicts with one's ideological commitments.
3. For simplicity, I call these beliefs biological facts in the text, but strictly speaking that is wrong, even if they were to supervene only on biological facts. The size of a buffalo herd supervenes on the individuals of the herd, but is not itself a property of individuals. But this just returns to us what must be Searle's original error, that the sleeping brain can be fully described neurophysiologically, at the ground level.
4. *Cf.* Simon's (1966, pp. 28–36) interesting discussion of this phenomenon.

   I have heard two attempted rebuttals to my interpretation here. (1) Perhaps the brain passes the problem off to a syntactic box which works it over until it is solved. But what is the magic border which strips the semantics off the problem when it goes in the box and restores it when it comes out? This is just an *ad hoc* redescription intended to save the supposed connection between semantics and consciousness. (2) Perhaps the brain passes the problem off to the periphery of consciousness; there is a trace of consciousness which accounts for the eventual solution of the problem. This again smacks of 'adhocness', especially considering that the background problem-solving may span a period of *days*—if you like, days which are filled with exciting, consciousness-occupying events. What kind of consciousness can it be if you are not aware of it *at all*?
5. Putnam goes a bit farther, saying it is absurd also to deny them intelligence (p. 12). That seems sensible enough. But for the purposes of this paper I want to treat semantics, intentionality, and intelligence in common.
6. I am not arguing that there is *no* conceptual relation between behaviour and intelligence, but just that there is no purely behavioural criterion which is alone sufficient for intelligence. Turing's substitute for the question 'Can machines think?', whatever its practical merits, as a *theoretical analysis* of machine intelligence leaves large holes for unintelligent robots to trundle through.
7. See Ian Hacking (1983), chapter 6.
8. There remains the possibility (rather, likelihood) that adopting the 'internal semantics' approach is both appropriate and necessary for current AI research. The lack of direct referential connections between knowledge representations and reality is no impediment to developing useful expert systems or conducting AI research in general.

   But Rapaport is going much further than recommending a certain methodological approach for AI. Not only is referential semantics 'not the kind of semantics that is of computational interest' (p. 84), which is the point above, but also 'this [referential] link is [not] necessary for *understanding natural language*' (p. 86). But, as I have indicated, reference is an essential component of natural language; it can be eliminated only at the cost of denaturing language. Rapaport has a variety of interesting arguments in defence of an unworldly semantics for artificial intelligence, but giving these their due would take us too far afield.
9. The following argument is from Putnam (1981), chapter 2 (for a more general argument see his appendix). Take the sentence (1) A cat is on a mat. This can be reinterpreted to assert (of this world) that a cherry is on a tree while preserving the truth and falsity of any set of English sentences. Consider the three sentences:

   (a) Some cat is on some mat and some cherry is on some tree.
   (b) Some cat is on some mat and no cherry is on any tree.
   (c) Neither (a) nor (b) are true.

   Then define cat* and mat*:

   $x$ is a cat* iff (a) i
   $x$ is a cherry.

   $x$ is a mat* iff (a)
   $x$ is a quark.

   Then there are three class
   mat* is true if and only if

   (i) Worlds where (a) i
   some tree.
   (ii) Worlds where (b)
   world what (1) ass
   (iii) Worlds where (c)
   have to be true (b
   (2) is also false, si

   Since our world falls into c
   world) about cherries and
   set of sentences.
10. It is a recent but important
    is not 'all in the head,' that
    syntax *is* all in the head, sy
    in, for example, Burge (19
11. A defence of Searle's ans
    understanding. A more di
    whether he can extend his
    man who has visited outer
    or astronaut for that matter
    to say 'No,' since he is just
    no provision for inventing
    astray, for, however rote t
    Test, and that means it mu
    The language rules must b
    of his Chinese room.
12. I am indebted to John Win
13. Note that Winnie's example
    derived from iconic represe
    There is no reason at all w
    agreeable. (*Cf.* Harnad's T

    On another point, I beg
    to resolve the symbol-ground
    between the symbols and t
    principle why, say, cameras
    from which 'invariants' can
    That is, there is nothing in
    digital converters from play
    symbol system: the camera
    matter, bit-mapped iconic in
    symbols in a language.) But
    to interact with the world a
    nonsymbolic subsystems, w
    just a non-starter. (To be s
    dealing with the dispute bet
14. This paper was presented a
    Art Form?' at the SUNY Bi
    June 21–23, 1990. I am grate
    Koertge, John Winnie and J

## References

Burge, T. (1979) Individualism a
University of Minnesota.
Churchland, P. and Churchland
32–37.
Feyerabend, P. (1975) *Against M*

$x$ is a cat* iff (a) is true and $x$ is a cherry, or (b) is true and $x$ is a cat, or (c) is true and $x$ is a cherry.

$x$ is a mat* iff (a) is true and $x$ is a tree, or (b) is true and $x$ is a mat, or (c) is true and $x$ is a quark.

Then there are three classes of possible world to consider and in all of them (2) A cat* is on a mat* is true if and only if (1) is true:

(i) Worlds where (a) is true. Then (1) is trivially true. (2) is true because some cherry is on some tree.

(ii) Worlds where (b) is true. Then (1) is trivially true. (2) is true, because it asserts of this world what (1) asserts.

(iii) Worlds where (c) is true. Then (1) is false, because otherwise one of (a) or (b) would have to be true (by the law of excluded middle on 'some cherry is in some tree'). But (2) is also false, since no cherry is on a quark in any possible world.

Since our world falls into category (i), it follows that we can interpret (1) as an assertion (of our world) about cherries and trees without sacrificing any of the truth values we ascribe to any finite set of sentences.

10. It is a recent but important theme in philosophy—stemming from Putnam (1975)—that semantics is not 'all in the head,' that meaning is sensitive to social, linguistic context. Since our (biological) syntax *is* all in the head, syntacticism runs a foul of Putnam's good arguments and their relatives in, for example, Burge (1979).
11. A defence of Searle's answer could be attempted which has nothing to do with feelings of understanding. A more discriminating test of Searle's Chinese understanding would be to ask whether he can extend his linguistic abilities appropriately. Confronted for the first time with a man who has visited outer space, it is no great feat for any of us to invent the term spaceman—or astronaut for that matter. Can the Searle-Room do the same? Perhaps Searle would be inclined to say 'No,' since he is just following 'rote' rules for Chinese-token manipulation and has made no provision for inventing new squiggles or squoggles. But that just shows how intuitions *can* go astray, for, however rote the rules, the Chinese system must be capable of passing the Turing Test, and that means it must be capable of inventing new and appropriate symbols on demand. The language rules must be extensible and revisable, in contrast to Searle's spartan description of his Chinese room.
12. I am indebted to John Winnie for suggesting this example.
13. Note that Winnie's example no longer does any damage. The symbols representing my shirt are derived from iconic representations, and these can reflect perfectly well the colour of my shirt. There is no reason at all why Roboman couldn't tell me the colour of my shirt, so long as he is agreeable. (*Cf.* Harnad's Total Turing Test in (1989*b*); see also Rey 1986.)

    On another point, I beg to differ with Harnad that traditional AI doesn't have the resources to resolve the symbol-grounding problem. He says 'In a pure symbolic model the crucial connection between the symbols and their referents is missing...' (1989a, p. 17). But there is no reason in principle why, say, cameras yielding *digital* images can't be a source of iconic visual representations from which 'invariants' can be abstracted, producing categorical representations, and so forth. That is, there is nothing in Harnad's symbol-grounding story which excludes standard analog-to-digital converters from playing the symbol-grounding role. Of course, this is no longer a *pure* symbol system: the camera, for example, is not purely symbolic (e.g. it has lenses). (For that matter, bit-mapped iconic images 'downstream' from the camera, while being digital, are also not symbols in a language.) But *any* robot must have some kind of nonsymbolic transducer in order to interact with the world at all. Those people who ever took strong AI to disallow the use of nonsymbolic subsystems, whether advocates or opponents, were arguing about a thesis that is just a non-starter. (To be sure, Harnad 1989a is playing a somewhat different game, namely, dealing with the dispute between connectionism and traditional AI.)
14. This paper was presented at the workshop 'Artificial Intelligence: Emerging Science or Dying Art Form?' at the SUNY Binghamton Program in Philosophy and Computer & Systems Science, June 21–23, 1990. I am grateful for discussions with participants and with Tim van Gelder, Noretta Koertge, John Winnie and John McEneaney.

## References

Burge, T. (1979) Individualism and the mental, *Midwest Studies in Philosophy*, **4**: 73–121. Minneapolis: University of Minnesota.

Churchland, P. and Churchland, P. S. (1990) Could a Machine Think? *Scientific American*, **262**: 32–37.

Feyerabend, P. (1975) *Against Method* (London: Verso).

Hacking, I. (1983) *Representing and Intervening* (New York: Cambridge University Press).
Harnad, S. (1987) Category induction and representation. In S. Harnad (ed.) *Categorical Perception: The Groundwork of Cognition* (New York: Cambridge University Press), pp. 535–565.
Harnad, S. (1989*a*) The Symbol Grounding Problem presented at the CNLS Conference on Emergent Computation. Unpublished manuscript (forthcoming in *Physica D*).
Harnad, S. (1989*b*) Minds, machines and Searle, *Journal of Experimental and Theoretical Artificial Intelligence*, **1**: 5–25.
Hofstadter, D. R. and Dennett, D. (1981) *The Mind's I* (New York: Basic Books).
Marcel, A. J. (1983) Conscious and unconscious perception, *Cognitive Psychology*, **15**: 197–237.
Marx, M. H. (1984) Information processing (unconscious). In R. J. Corsini (ed.) *Encyclopedia of Psychology* (New York: Wiley), Vol. 2, pp. 216–218.
Putnam, H. (1975) The Meaning of 'Meaning'. In K. Gunderson (ed.) *Language, Mind and Knowledge* (Minnesota Studies in the Philosophy of Science, VII. Minneapolis: University of Minnesota).
Putnam, H. (1981) *Reason, Truth and History* (New York: Cambridge University Press).
Pylyshyn, Z. (1984) *Computation and Cognition* (Cambridge, Massachusetts: MIT Press).
Rapaport, W. (1988) Syntactic semantics. In J. Fetzer (ed.) *Aspects of Artificial Intelligence* (Dordrecht, Kluwer Academic Publishers).
Rey, G. (1986) What's really going on in Searle's Chinese room, *Philosophical Studies*, **50**: 169–185.
Searle, J. (1981) Minds, Brains and Programs, *Behavioral and Brain Sciences*, **3**: 417–457.
Searle, J. (1984) *Minds, Brains and Science* (Cambridge, Massachusetts: Harvard University Press).
Searle, J. (1990) Is the brain's mind a computer program? *Scientific American*, **262**: 26–31.
Searle, J. (forthcoming) Consciousness, explanatory inversion and cognitive science, *Behavioral and Brain Sciences*. References are to the final draft for commentators.
Simon, H. (1966) Scientific discovery and the psychology of problem solving. In R. G. Colodny (ed.) *Mind and Cosmos* (Pittsburgh: University of Pittsburgh), pp. 22–40.

J. Expt. Theor. Artif. In

*Research Paper*

# Inductive r

G. KALKAN

*University of*
*(UMIST), De*

*Abstract.* The problem
major bottleneck in the d
approach to alleviate this
accept, as input, a set
and produce, as output
described by a set of att
A common characteristi
are empirical in nature
inductive rule generation
That performance is asse
data. This paper demons
the rule generation and
interrelated and should
generation and rule inte
performance of existing i

*Keywords:* induction, at
overlapping

*Received 1 July 1991; rev*

## 1. Introduction

In the attribute-based in
induction is to determine
from a set of training data
new non-training data to

Since the training data
inductive rule generation
of decision classes of the
belong to one decision c
subclusters of the inital cl
conceptual description, b
conceptual description (wh
of attribute values) is the
rule is the class of the insta
has been generated. The ba