2

RSRE
MEMORANDUM No. 4391

ROYAL SIGNALS & RADAR ESTABLISHMENT

ON UNDERSTANDING IN COMMUNICATING SYSTEMS AND THE ERROR IN THE CHINESE ROOM THOUGHT-EXPERIMENT

Authors: T A D White & Dr S C Giess

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
RSRE MALVERN,
WORCS.

RSRE MEMORANDUM No. 4391

CONDITIONS OF RELEASE

0088205 BR-115853

************************* DRIC U

************************* DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

Royal Signals and Radar Establishment

Memorandum 4391

Title: On Understanding in Communicating Systems and the Error in the Chinese Room Thought-Experiment

Authors: T A D White and Dr S C Giess

Date: November 1990

Abstract

John Searle has proposed the Chinese Room thought-experiment to refute the notion of what he calls strong AI (Artificial Intelligence), that is that thinking is merely the manipulation of formal symbols.

We discuss the Chinese Room thought-experiment, and show that Searle's argument is in error.

We show, therefore, that the Chinese Room thought-experiment provides no grounds for the conclusion that computer hardware, solely on the basis of running computer software, can never be regarded as intelligent.

THIS PAGE IS LEFT BLANK INTENTIONALLY

Contents

THIS PAGE IS LEFT BLANK INTENTIONALLY

1 Introduction

1.1 Background and Aim

John Searle has proposed the Chinese Room thought-experiment [1, 2] to refute the notion of what he calls strong AI (Artificial Intelligence), that is that 'thinking is merely the manipulation of formal symbols' [2].

As a corollary, Searle proposes to refute the idea that symbol-manipulating devices can be regarded as intelligent by virtue of their manipulation of formal symbols, and, in particular, he proposes to refute the idea that digital computers can be regarded as intelligent solely on the basis of running computer programs.

We discuss the Chinese Room thought-experiment using thought-experiments of our own, and show that Searle's argument is in error.

We note that we are not claiming to have confirmed the notion of strong AI; what we are saying is that Searle's Chinese Room thought-experiment is not a refutation of it.

1.2 The Turing Test

At the heart of the Chinese Room thought-experiment lies the Turing test. The Turing test is accepted by many as the definitive scientific test for demonstrating understanding in a machine.

We have drawn a brief description of the Turing test from [2[a], 3, 4] for those who are unfamiliar with it. The Turing test is that if an observer cannot distinguish, on the basis of written answers to questions, between a person and a machine, then we say that the machine is demonstrating understanding.

We note that the observer questions two entities, and that the Test's result is a consequence of the observer being unable to distinguish between them; we note also that the person, as well as the machine, receives and answers questions.

We note this because we wish to contrast this *formulation* of the Turing test with Searle's *use* of it. We do this later in Section 1.4; but first, we describe the Chinese Room thought-experiment.

[a] Reference 2, Section 1, paragraph 3.

1.3 The Chinese Room, Thought-Experiment, Argument and Conclusion

Searle accepts that certain kinds of machine are, or may be constructed to be, intelligent [2].

He argues, however, that he has a refutation of strong AI: he has a machine which can pass the Turing test, but which, Searle claims, can be shown to lack understanding. The Chinese Room is the machine Searle has designed.

Searle's Chinese Room comprises:

> a set of formal symbols—that is, the characters of the Chinese language;
>
> a set of rules for manipulating the symbols—that is, rules for the manipulation of Chinese characters; and
>
> a mechanism which does nothing more than obeys the given rules—that is, some mechanism which, by hypothesis, has no understanding of the Chinese language *(indeed, in Searle's descriptions* [1, 2], the mechanism is the non-Chinese-speaking Searle himself).

For completeness, we note that the mechanism and rules are contained within a room, that the symbols can be recognized entirely by their shape (this is what Searle means by formal), and that the mechanism has some means of input and output through the walls of the room.

Searle's thought-experiment is a Turing test of the Chinese Room by a native Chinese-speaker.[b] The outcome of the thought-experiment is that the Chinese-speaker confirms that the output from the Room is sensible Chinese; thus the Room passes the Turing test for understanding Chinese.

Searle's argument is that the Room cannot understand Chinese for, while its output is consistently sensible, we know that it is merely a mechanism obeying rules for symbol manipulation. Searle justifies this by observing that, after conducting the thought-experiment, he, the mechanism, has gained no appreciation of the Chinese language and therefore could not have understood Chinese during the thought-experiment.

[b] There is a second, and a most important, hypothesis here: the native Chinese-speaker understands Chinese. Checking this second hypothesis will reveal some interesting results.

Searle's conclusion is that since digital computers, running computer programs, are precisely this kind of machine then computer hardware, solely on the basis of running computer software, can never be regarded as intelligent.

Searle also draws the corollary that, since the human mind is intelligent, it cannot work on this principle of brain-hardware running mind-software.

1.4 Closing Words—Contrasts and Consequences

When Searle uses the Turing test in the Chinese Room thought-experiment only the machine, the Chinese Room, receives and answers questions. Moreover, the questions are posed, not by an observer, but by the other entity in the thought-experiment, the person, a native Chinese-speaker.

We note that the native Chinese-speaker is not questioned (there is no-one to pose the questions) and we conclude, therefore, that the assumption that the native Chinese-speaker understands is brought into Searle's thought-experiment and forgotten (see footnote b).

We believe that establishing the credentials of both participants in the Test, by an observer posing questions, is a most important ingredient of the Turing test. We believe that Searle's failure to maintain symmetry results in a weaker form of the Test.

As a consequence, we conduct our own thought-experiments in this style: we have an observer who poses questions to the Chinese Room *and* to the native Chinese-speaker.

THIS PAGE IS LEFT BLANK INTENTIONALLY

2 Preparing for New Experiments

2.1 Recognizing Confusion

Our impression of the Chinese Room thought-experiment is that it is not well formulated in spite of, or perhaps because of, Searle's graphic accounts. For example, we see in [1]:

> ' ... my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody ... can tell that I don't speak a word of Chinese.'

What are we to make of this? We think that 'I' is used in two senses, two senses which are not yet proven the same. We think that what is meant is:

> ' ... my (the Chinese Room's) answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody ... can tell that I (Searle without the rules) don't speak a word of Chinese.'

As a second example, we see in [2]:

> ' ... I satisfy the Turing test for understanding Chinese. All the same, I am totally ignorant of Chinese.'

Again in this passage, 'I' is used in two different senses which are not yet proven the same. Our best interpretation of the passage is:

> ' ... I (the Chinese Room) satisfy the Turing test for understanding Chinese. All the same, I (Searle without the rules) am totally ignorant of Chinese.'

We are no pedants when we try to understand this clearly, for these switches of meaning of 'I' represent less than convincing evidence for Searle's argument. Indeed, this is not evidence: it is the very point of debate.

2.2 Clearing Confusion—The Players

Let us be clear about the players in the Chinese Room thought-experiment, they will be players in our thought-experiments also. They are the Chinese Room, a Chinaman, the Observer and the Protagonist.

<u>The Chinese Room</u> The Chinese Room comprises a mechanism, Chinese characters and rules as described in Section 1.3. The Chinese Room passes the Turing test.

The Chinaman The Chinaman is a native Chinese-speaker who converses with the Room and is unaware of its construction. (We note that the phrase 'native Chinese-speaker' is cumbersome but that the simpler 'Chinese' can mean either a native of China or the Chinese language. We shall refer to the native Chinese-speaker as 'the Chinaman', an archaic word which, in other circumstances, we would avoid.)

The Observer We have personified the audience to whom our discussion is addressed. We assume that the Observer speaks English to make inquiries of the Chinaman and of the Chinese Room. We assume also that the Observer speaks Chinese to make further inquiries. Indeed, for reasons that will become clear (see Section 3.2), we assume the Observer speaks all languages of the world.

The Protagonist We have personified, separately, a part of the audience: the cause the Protagonist champions is Searle's Chinese Room argument. The Protagonist speaks English only.

When discussing Searle's thought-experiment, we need to be clear about the role in which Searle debates. Thus, when we think Searle is presenting his argument we use the term 'Searle-the-Protagonist', when Searle is the Room's mechanism we use the term 'Searle-the-mechanism', and later, when Searle internalizes the rule-book used by the mechanism, we use the term 'Searle-the-Room'. We trust these slightly awkward phrases will benefit overall clarity.

2.3 Clearing Confusion—A Rigorous Form of Searle's Result

Let us be quite clear about Searle's result. That is, let us be quite clear about what happens after the successful Turing test of the Chinese Room.

What happens is that Searle-the-mechanism asserts, from within the Room[c] and in English, 'I am totally ignorant of Chinese ... ' [2]. As we have noted, Searle is far from straightforward in his use of the word 'I', and this result needs to be made precise before we can proceed. We re-cast Searle's result into a rigorous form of question-and-answer similar to that of the Chinese Room thought-experiment itself.

[c] It must be from within the Chinese Room. There is no argument that Searle-the-mechanism, separated from the rules and outside the Chinese Room, is ignorant of Chinese. The argument is whether Searle-the-mechanism, with the rules and inside the Room, is ignorant of Chinese.

We suggest that what happens is that Searle-the-Protagonist conducts another test on the Chinese Room: he asks a question and receives a reply. From outside the Room, Searle-the-Protagonist asks Searle-the-mechanism, inside the Room, a question, in English, to the effect:

> 'what do you remember of your recent Chinese conversation?'

Searle's result is that the response will be something like:

> 'nothing except that when I saw such-and-such Chinese characters I replied with such-and-such other Chinese characters.'

That is, the reply will be in terms of the syntactic manipulation of Chinese characters. Searle's argument is that now it has been shown that the Room did not understand Chinese during the Turing test at all.

We have no dispute with the result: well, actually we do. What we mean is, on the one hand, that we must accept that Searle asserts what he asserts, but, on the other hand, that we can argue, and do argue, about the validity and meaning of the assertion. In our rigorous form of Searle's result we mean, on the one hand, that we must accept that Searle-the-Protagonist conducts a second test, asking a question and receiving an answer, but, on the other hand, that we argue about the validity of the asking and about the meaning of the reply.

Therefore, we can dispute, and do dispute, Searle's argument and conclusion.

2.4 Clearing Confusion—Rigorous Experiments

In order for us to be convinced that the Protagonist's second test on the Chinese Room is scientifically meaningful, we have to compare rigorously the Room with the Chinaman. In other words, we must conduct the second test on the Chinaman as well as conducting the test on the Room.

This is what we do in the three W-G thought-experiments following.

THIS PAGE IS LEFT BLANK INTENTIONALLY

3 Three Thought-Experiments

3.1 The Starting Point

The starting point is that the Chinese Room has been successful in a Turing test for understanding Chinese which has been conducted by a Chinaman, and that after completion of the Test, Searle, in the role of the Room's mechanism, protests in English 'I satisfy the Turing test for understanding Chinese. All the same, I am totally ignorant of Chinese ... ' [2].

We introduce some notation: throughout the descriptions of our experiments, English dialogue is represented in the Times font, Chinese in the Avont Gorde font, and an arbitrary language in the Courier font; also, when describing dictionaries, we write 'Chinese-to-English' when we mean Chinese-English:English-Chinese.

Now we conduct our thought-experiments. We conduct further question-and-answer experiments on the Chinese Room, not forgetting to compare the results with similar experiments on the Chinaman.

3.2 The First W-G Thought-Experiment

In the first W-G thought-experiment, we reveal a methodological error in Searle's approach. The first experiment is described in detail in Annex A.

We ask the Chinaman the same English question we posed to the Chinese Room in Section 2.3:

> 'what do you remember of your recent Chinese conversation?'

No reply is forthcoming. Rather than conclude that the Chinaman does not understand Chinese, we teach the Chinaman English. Then, to maintain equality, we teach the Room English too, and in exactly the same way.[d] Having done this, the Chinaman answers our question in English, and the Room no longer replies in terms of syntactic manipulation but gives a satisfactory, English account of the conversation. The Chinese Room understands the recent Chinese conversation.

We generalize the first experiment to show that the use of English, in particular, is irrelevant.

[d] It is important that we teach the Room in exactly the same way. Not only does this preserve equality of treatment, but also it guarantees success: for otherwise, we would have a test whereby to distinguish the Room from the Chinaman.

An Observer may use any language and the result of the first experiment is the same: given equal treatment of the Chinese Room and the Chinaman, it is impossible simultaneously to confirm that the Chinaman understands Chinese while failing the Room as not understanding.

The conclusion from our first experiment is that Searle has failed to treat equally the Chinese Room and the Chinaman. Searle has failed, therefore, to ensure the conditions under which English is a valid vehicle for communication and Searle's English assertion is meaningless.

3.3 The Second W-G Thought-Experiment

In the second W-G thought-experiment, we reveal why Searle can utter his assertion at all: it follows from a design error in the Chinese Room. The second experiment is described in detail in Annex B.

We ask our English question, not of the Chinaman nor of the Chinese Room, but of the Chinese Room's mechanism. This involves making rigorous the notion of addressing the mechanism directly, for the mechanism, when inside the Room, normally communicates in Chinese characters only. Two things emerge.

First, **assuming** we can access the mechanism directly, a result similar to that of our first experiment is produced. That is, whatever language the mechanism uses, it is impossible simultaneously to confirm that the Chinaman understands Chinese while failing the mechanism as not understanding.

Secondly, by virtue of the mechanism becoming accessible, the Chinese Room is no longer a Chinese Room: our use of the word 'normally' two paragraphs ago is woefully inadequate, 'always' is the correct word. A Chinese Room's mechanism cannot assert 'I am totally ignorant of Chinese'—at least, not in English. In contrast, the mechanism of a Chinese-Room-where-there-is-direct-access-to-the-mechanism can; but then the statement is not a comment on the mechanism's ability in Chinese, it is a statement on its inability to translate between Chinese (the subject of the assertion) and English (the vehicle in which the assertion is made).

Searle wished, we assume, to design the Chinese Room to demonstrate a lack of understanding of Chinese. We conclude that he has made a design error and has constructed, instead, a Room which has single-language ability in Chinese and single-language ability in English. The Room's only lack is an ability to translate between the two. We explore this phenomenon in our third experiment.

3.4 The Third W-G Thought-Experiment

In the third W-G thought-experiment, we confirm the effect of the design error by providing a Chinese-to-English foreign language dictionary to the Chinaman and to the Chinese Room *inter alia*. The third experiment is described in detail in Annex C.

We show that the supply of a foreign language dictionary by itself does not give the Chinaman fluent ability in English: we show, however, that the supply of a dictionary is significant in the case of Searle's design-flawed Chinese Room. We show why our first and second experiments produce the results they do.

We conclude that Searle's design error is allowing direct access to the mechanism of the Room: this introduces the idea that the Chinese Room has English as a second language,[e] whereas the idea is actually frustrated by Searle's exclusion of a Chinese-to-English dictionary.

3.5 Conclusion

We have identified two errors in Searle's Chinese Room thought-experiment. They are one of methodology, that is failing to continue to compare the Chinese Room with the Chinaman, and one of design, that is allowing direct access to the Room's mechanism.

Thereby, Searle fails to ensure the conditions under which English is a valid vehicle for communication and the utterance 'I am totally ignorant of Chinese ... ' is meaningless. Indeed, it cannot be uttered at all by other than a design-flawed Chinese Room; and with the design flaw Searle introduces the idea that the Chinese Room has a second language (see footnote e) when he has ensured that it does not.

We conclude that these errors are the instruments Searle uses in his argument: a one-sided 'comparison', and the English language which 'understands' nothing of Chinese.

We reject Searle's conclusion that the Chinese Room cannot have understood Chinese during the Turing test as an illusive consequence of these errors.

[e] In the way people use the term of people: that is, a first language, a second language and an ability to translate between them.

THIS PAGE IS LEFT BLANK INTENTIONALLY

4 Closing Words

4.1 Syntax or Semantics

We have come to appreciate the power of the mere syntactic rules used in the Chinese Room. The rules allow the Room to be accepted as a native Chinese-speaker. The rules governing the symbol 'chair', say, must cover all legitimate uses of the symbol 'chair', and they must forbid all unacceptable uses. We find it difficult to avoid concluding that such rules contain all that we might call the semantics of 'chair'.

4.2 Semantics and Reality

Also, we have come to appreciate the strangeness of Searle's set-up, flawed though the method and the design are. Searle has designed into the Chinese Room a separation which we do not expect in the world-at-large when we are talking about people with two natural languages.

The effect is odd since we are unused to there being no connection between natural languages such as Chinese and English. We take this for granted and, indeed, examples of unconnected, undeciphered natural languages are rare. Moreover, for these languages there is some hope of translation yet: we live in the space where natural languages arise and which allows the connection to be made. The relationship between natural languages relies on some intermediary, say, human sensory reality; this is the 'dictionary' between natural languages which allows us to create the books we call dictionaries.

We labour this point for two reasons. The first is that we cannot see any intrinsic reason for languages to be translatable into one another; languages, in principle, can exist in splendid isolation. For example, we can imagine a message being beamed at us from beings who may share a very different sensory reality from our own subjective terrestrial one; we can now ask ourselves: what are the exact conditions necessary to understand the message? The second reason is that if we deny the existence of an objectively correct reality and prefer the belief in a subjective personal one, then our reality is derived from our senses. Now, sound waves, light waves, nerve impulses and the like are all formal symbols in the sense that is used in Searle's Chinese Room thought-experiment; so we may ask: is there any reason to assume that we are different from symbol-manipulating machines? We hope to expand upon these ideas in a later paper.

References

[1] Searle J R, Minds, brains and programs, The Behavioural and Brain Sciences (1980), 3.

[2] Searle J R, Is the Brain's Mind a Computer Program, Scientific American, January 1990.

[3] Hodges A, Alan Turing: The Enigma, Burnett Books, 1983.

[4] Turing A M, Computing Machinery and Intelligence, Mind, Vol LIX, No 236 (1950) *excerpt in* Hofstadter D R and Dennett D C, The Mind's I, Penguin, 1982.

Annex A—The First W-G Thought-Experiment

A.1 Motivation

The motivation for this experiment is two-fold.

First, while reading Searle's English descriptions of the Chinese Room thought-experiment, motivation developed from a growing puzzlement about the relevance of English to understanding Chinese. We suspect that there are many million single-language native Chinese who are quite content that they understand the Chinese language without any reference whatsoever to English. We explore this further in the second and third W-G experiments.

Secondly, motivation developed from the realization that, after the Turing test of the Chinese Room, Searle abandons any further consideration of the Chinaman. Our argument here is simple: whatever test we use to test the Room, we must use the same test to test the Chinaman; and moreover that test must show that the Chinaman understands Chinese. Our reasoning is this: if a test of the Chinaman cannot show that the Chinaman understands Chinese, why then there's something wrong with the test; and if we do not apply the same test to both the Room and the Chinaman, why then we could prove anything.[f]

The first W-G thought-experiment investigates when it is valid to use English as a vehicle to communicate with the Chinese Room.

A.2 The Experiment

The Observer poses the following question, in English, to the Chinaman:

> 'what do you remember of your recent Chinese conversation?'

The Chinaman, speaking no English, will not be able to answer at all. The Observer is careful not to conclude that the Chinaman does not understand Chinese. Instead, the Observer teaches the Chinaman English. This means that the Observer gives the Chinaman:

> first, the rules for manipulation of English words in Chinese;[g] and

[f] As an illustration of this we offer the following sketch: Salesman to Customer, 'Our product is more robust than our competitor's. See, when I hit theirs with a hammer, it breaks.' In these circumstances, the wise customer asks for the salesman's product to be hit with a hammer too.

[g] We can do this by assumption of Searle's paradigm. But let us imagine we give the Chinaman one of those speed-teaching language courses often advertised in the newspapers (Chinese newspapers! for English newspapers advertise language courses for English-speakers). Such a course is arguably a set of rules and a dictionary.

secondly, a Chinese-to-English foreign language dictionary which contains rules for mapping Chinese characters into English words and *vice versa.*

The Chinaman, having learnt English, now replies, in English, with a description of the recent conversation. The Observer has confirmed the hypothesis that the Chinaman understands Chinese, using English as the vehicle.

But—and this is the point—for the Observer to pose the question to the Chinaman and for the Observer to receive the reply that hypothesis demanded, the Observer had to supply something to the Chinaman. Scientific rigour says that whatever we do to one entity in a test of companson, we must do to the other. So now the Observer must supply the Chinese Room with:

the rules for manipulation of English words in Chinese; and

a Chinese-to-English dictionary.

Now, the Chinese-to-English dictionary results in the Room being able to reply, in English, with a description of the recent conversation. The reply is, of course, good English since the rules for manipulation of English words ensure this is so.

In other words, teaching English to the Chinaman allows the Chinaman to reveal an understanding where previously the response was far from promising; and teaching the Room English, in an identical way as a foreign language,[h] allows the Room to reveal its understanding where previously it had been suggested, by the Protagonist, that the Room had none.

The result of our first experiment is that both the Chinaman and the Chinese Room can give the Observer a satisfactory account, in English, of their understanding of the recent Chinese conversation.

A.3 A Generalization

The language the Observer uses to pose the question is quite irrelevant. The question:

[h] It is important that we teach the Room in an identical way. Not only does this preserve equality of treatment, but also it guarantees success: for otherwise, we would have a test whereby to distinguish the Room from the Chinaman.

```
'what do you remember of your recent Chinese conversation?'
```

can be posed in any language.

Posing this question is an experiment, and it must be carried out on the Chinaman also. To do this, the Observer would teach the Chinaman the necessary language as a foreign language.[i] Then, to remain scientifically valid in the comparison, the Observer must teach the Room the same language, similarly as a foreign language. Both the Chinaman and the Chinese Room would be able to give a satisfactory account of their understanding of the recent conversation.

The result of this generalized form of our first experiment is that the Observer, using no language, can simultaneously confirm that the Chinaman understands Chinese and fail the Chinese Room as not understanding.

A.4 Conclusion

We conclude that Searle has failed to preserve equality of treatment of the Chinaman and the Chinese Room; thus, he has failed to ensure the conditions under which English is a valid vehicle for communication.

[i] The exception is if the Chinese language is chosen, for then we having nothing to do.

THIS PAGE IS LEFT BLANK INTENTIONALLY

Annex B—The Second W-G Thought-Experiment

B.1 Motivation

We posed our question to the Chinese Room, whereas in Searle's original thought-experiment it is Searle-the-mechanism who asserts 'I am totally ignorant of Chinese ... '. The second W-G thought-experiment establishes what has to be done to ask mechanisms questions.

B.2 There Are Two Problems ...

There are two problems in posing questions to the Room's mechanism. First, it is not clear how the Observer can talk directly to the mechanism while it is still part of the Chinese Room; secondly, it is not clear what the equivalent experiment on the Chinaman is. These problems are solved later (see Section B.4). We deal now with posing questions directly to the mechanism of the Chinese Room, accepting that we are assuming much. So, **assume** the Observer can address the Room's mechanism directly.

B.3 The First Part of the Experiment

Let us say the mechanism's language is *MC*. For the Observer to pose questions in English, the Observer would have to teach English to the *MC*-speaking mechanism. Then, to preserve equality of treatment, the Observer would 'teach' English to the Chinaman, and in exactly the same way. We put 'teach' in quotes since the teaching is pointless: the Chinaman doesn't understand *MC* so cannot learn English that way. But this is no matter.

The mechanism now speaks English and we invoke our first experiment.

> We pose our English question to the mechanism and to the Chinaman only to discover no reply from the Chinaman. Since we have to confirm the hypothesis that the Chinaman understands Chinese, the next step is to teach the Chinaman English.
>
> Then, to preserve equality of treatment, we teach English to the mechanism also, and in exactly the same way. This appears to be a wasted exercise because the mechanism has already been taught English. But—this is the point—it is not a wasted exercise: previously the mechanism was taught English as being foreign-to-*MC* and an *MC*-to-English foreign language dictionary was provided; now the mechanism is taught English as foreign-to-Chinese and a Chinese-to-English dictionary is provided.
>
> We invoke the argument of the first W-G experiment to show that a satisfactory

understanding of the recent Chinese conversation is revealed both by the Chinaman and by the mechanism.

But this does not depend on the mechanism's language. The result is, whatever the mechanism's languag 'he Observer cannot simultaneously confirm that the Chinaman understands Chinese anu fail the mechanism as not understanding.

B.3.1 Searle's Assertion

The argument applies not only to our rigorous form of question-and-answer but also to Searle's original description; here Searle-the-mechanism asserts, from within the Room and in English, 'I am totally ignorant of Chinese ... '

The Observer can confirm, in some language, the hypothesis that the Chinaman understands Chinese. Our first experiment deals with the Observer's language; the first part of our second experiment deals with the mechanism's language. The Observer can then further test the mechanism, with its rules, only to discover that a satisfactory account of the conversation is forthcoming from the mechanism. The mechanism protests its ignorance and yet is able to demonstrate an understanding of the conversation. This seeming paradox is explained later (see Sections B.4.2 and B.4.3 and also Annex C).

B.3.2 The Mechanism's Construction

Finally, before proceeding, we note that we have made no assumptions about the construction of the mechanism: the mechanism could be an *MC* Room. Indeed, in Searle's description, the Observer cannot tell whether Searle-the-mechanism is a native English-speaker or an English Room of Turing test quality. For this reason, if we accept Searle's conclusion that the Chinese Room doesn't understand Chinese, then we must reserve judgement about Searle-the-mechanism's understanding of English.

B.4 The Second Part of the Experiment

In the first part of the experiment we assumed access to the mechanism; now we arrange such access completing our ideas of what has to be done to ask mechanisms questions. Following leads given by Searle himself, we consider two cases.

First, the walls of a Chinese Room are removed allowing the Observer direct access to the mechanism, which remains of unknown construction. The language the mechanism uses is English and it continues to consult an external rule-book for Chinese character manipulation.

We call this the Exposed Chinese Room.

Secondly, the walls of a Chinese Room are removed as before. The mechanism is also as before, except that the rules of manipulation are incorporated internally within it. We call this the Open Chinese Room, and when Searle plays this role we use the term 'Searle-the-Room'.

And now we have solved the two problems we described earlier in Section B.2. First, we have arranged direct access to the mechanism (in two different ways) so our experiment is legitimate. Secondly, since we have direct access, the mechanism is not a hidden internal detail, and we need not concern ourselves with what the equivalent hidden internal detail of the Chinaman might be. We can conduct our experiment on the Chinaman equivalently.

The Observer is equipped with a Chinaman, a Chinese Room, an Exposed Chinese Room, and an Open Chinese Room. The Observer conducts the second part of the experiment which involves posing the following four questions to each of these four entities:

in Chinese,

'do you understand the Chinese questions put to you?'; (1)

'do you understand English questions if put to you?'; (2)

and in English,

'do you understand English questions if put to you?'; (3)

'do you understand the Chinese questions put to you?'. (4)

We suggest that the following responses are obtained:

Question	The Chinaman	Chinese Room	Exposed Chinese Room	Open Chinese Room
(1)	Yes	Yes	Yes	Yes
(2)	No	No	No	No
(3)	*can't answer*	*can't answer*	Yes	Yes
(4)	*can't answer*	*can't answer*	No	No

B.4.1 The Chinaman's Responses and the Chinese Room's Responses

The Chinaman's responses are self-evident. The Chinese Room's are identical; but they need explanation as they are different from those of the Exposed, and of the Open, Chinese Rooms.

Let us follow what happens when the third and fourth questions are put (in the case where the English-speaking Searle has taken the role of the mechanism).

> Formal symbols, English words, arrive in the Chinese Room. The mechanism, being conscientious, attempts to handle the symbols using the rules of manipulation. The mechanism fails, of course, since English words are not accommodated anywhere in the rules for manipulation of Chinese characters. Following the dictates of the rules, the mechanism issues some standard reply, in Chinese, 'pardon, I don't understand'.
>
> All this would happen to the intense frustration of the English-speaking Searle-the-mechanism, who can recognize the formal symbols but who cannot subvert the given rules. Quite simply, it is not in the Chinese language to answer non-Chinese questions and the mechanism cannot make it so.

B.4.2 The Exposed Chinese Room's Responses

The Exposed Chinese Room is different since we have ensured, by construction, that the Observer can talk directly to the mechanism. This means—and this is the point—that the mechanism can decide whether to invoke the external rules of manipulation or to answer directly. We have arranged that the language used by the mechanism of our Exposed Chinese Room is English, and hence arranged that our Questions 3 and 4 are likewise in English.

An Exposed Chinese Room, wherein the mechanism can be addressed directly, becomes by virtue of this no longer a Chinese Room, with ability in one language, but some sort of Chinese-and-English Room, with some sort of ability in two. It is by this means that Searle-the-mechanism, in Searle's graphic description, can assert, directly to the Observer and in English, 'I am totally ignorant of Chinese ... '.

We are suspicious of the 'some sort' of ability the Exposed Chinese Room has. We have made our suspicion clear in the form of Questions 2 and 4. Let us follow what happens when the four questions are put.

> Formal symbols, Chinese characters, arrive .at the mechanism. The mechanism recognizes them and decides that obeying the rules is appropriate. The rules allow the Chinese questions to be answered indistinguishably from a Chinaman. Just so, but what does the answer to Question 2 mean?
>
> Formal symbols, English words, arrive at the mechanism. The mechanism recognizes

> them and decides that direct answering is appropriate. The mechanism, being the English-speaking Searle-the-mechanism answers fluently. Just so, but what does the answer to Question 4 mean?

We believe we can explain and we have confidence in our explanation since we test it beforehand in our third experiment. But first, we describe the answers of the Open Chinese Room ... with apologies to those who by now have seen our line of reasoning.

B.4.3 The Open Chinese Room's Responses

The Open Chinese Room, too, can decide whether obeying the internalized rules is appropriate or whether answering directly is the correct course of action.

> Formal symbols, Chinese characters, arrive at the Open Chinese Room. The Open Chinese Room recognizes them and decides that obeying the internalized rules is appropriate. The rules allow the Chinese questions to be answered indistinguishably from a Chinaman. Just so, but what does the answer to Question 2 mean?
>
> Formal symbols, English words, arrive at the Open Chinese Room. The Open Chinese Room recognizes them and decides that answering in the mechanism's language is appropriate. The Room, being the English-speaking Searle-the-Room answers fluently. Just so, but what does the answer to Question 4 mean?

Again, we believe we can explain and we test our explanation in our third experiment.

B.5 Conclusion

We conclude that Searle has introduced a design error into the Chinese Room: he has constructed a Room which has single-language ability in Chinese and single-language ability in English. The Room's inability is not in Chinese nor English but in translation between the two.

THIS PAGE IS LEFT BLANK INTENTIONALLY

Annex C—The Third W-G Thought-Experiment

C.1 Motivation

It remains for us to explore Searle's so-called result.

Searle's result is that the mechanism of Chinese Room asserts 'I am totally ignorant of Chinese ... '. We have shown that the very ability to utter such an assertion, in English, is the result of a design flaw. But we have shown this at the expense of introducing a seeming paradox. The paradox is clear in Section B.3.1 where the mechanism protests its ignorance of Chinese, but, when questioned validly and equally with the Chinaman, can give an account of its understanding of Chinese. The paradox is made rigorous, in the form of questions and answers, in Questions 2 and 4 of our second experiment.

The third W-G thought-experiment examines what happens if only *partial* language capabilities are supplied.

C.2 The Experiment

The Observer supplies a Chinese-to-English foreign language dictionary—just the dictionary—to each of: a Chinaman; a Closed[j] Chinese Room, where it is held externally to the Room's mechanism; an Exposed Chinese Room,[k] where it is held externally to the directly-accessible, English-speaking mechanism; and an Open Chinese Room, where it is 'internalized' by the directly-accessible, English-speaking mechanism.

The Observer poses the following two questions to each of these four entities:

in Chinese,

'do you understand English questions if put to you?'; (2)

and in English,

'do you understand the Chinese questions put to you?'. (4)

We suggest that the following responses are obtained:

[j] We now call a Chinese Room which provides no access to the mechanism, a Closed Chinese Room.

[k] Recall that we have shown that Searle's Chinese Room, which has direct access to the mechanism, is either an Exposed Chinese Room or an Open Chinese Room.

Question	The Chinaman	Closed Chinese Room	Exposed Chinese Room	Open Chinese Room
(2)	No	No	Yes	Yes
(4)	*can't answer*	*can't answer*	Yes	Yes

C.2.1 The Chinaman's Responses and the Closed Chinese Room's Responses

The Chinaman's responses are self-evident. The supply of a mere foreign language dictionary does not convince the Chinaman that English questions could be understood with an English-like fluency, and, indeed, the Chinaman would not be able to handle fluently the English question, Question 4. The responses of the Closed Chinese Room are identical to those of the Chinaman. Let us follow what happens when the second and fourth questions are put (in the case where the English-speaking Searle has taken the role of the mechanism).

> Question (2) Formal symbols, Chinese characters, arrive in the Closed Chinese Room. Now, not only can the mechanism correctly manipulate these characters but also the mechanism can map them into other formal symbols, English words, using the external foreign language dictionary. But the rules say nothing about the manipulation of English words, and the mechanism cannot subvert them. At best a Chinese-structured answer, as dictated by the rules, could be given but with Chinese characters replaced by English words. This is the analogue of the Chinaman's lack of an English-like fluency.

> Question (4) Formal symbols, English words, arrive in the Closed Chinese Room. The external rules of manipulation do not cater for English words, but the English words can be mapped into Chinese characters using the external foreign language dictionary. Probably no sense could be made of the ensuing English-structured question where Chinese characters now replace the English words. This is the analogue of the Chinaman not being able to answer.

C.2.2 The Exposed Chinese Room's Responses

The responses of the Exposed Chinese Room are different.

> Question (2) Formal symbols, Chinese characters, arrive at the Exposed Chinese Room. The mechanism recognizes them and decides that obeying the external rules is appropriate. But also, the symbols can be mapped into English words and, since direct access to the mechanism is allowed, the English-speaking mechanism can make good sense of the Chinese-structured English words. The mechanism can make very good

sense indeed, because it is the English-speaking Searle-the-mechanism. The reply, translated back using the external dictionary, will be, by definition, fluent Chinese.

Question (4) Formal symbols, English words, arrive at the Exposed Chinese Room. The mechanism recognizes them and decides that direct reply is appropriate. But also, the Exposed Chinese Room can map the symbols into Chinese characters and, since rules of manipulation of Chinese characters are available, the English-speaking mechanism can make good sense of the English-structured Chinese. The reply will be, by assumption, fluent English.

C.2.3 The Open Chinese Room's Responses

The responses of the Open Chinese Room are identical to those of the Exposed Chinese Room and for the same reasons, except that the two sets of rules are held internally.

C.3 Conclusion

We conclude, unsurprisingly, that a foreign language dictionary, by itself, does not give full foreign-language ability. However, we conclude that augmenting rules of manipulation with foreign language dictionaries is significant when ability in more than one language is required. Indeed, this is why preserving equality of treatment in our first and second experiments produced the results it did: the Chinese Room, and later the mechanism, were taught English as foreign-to-Chinese and were so provided with the necessary Chinese-to-English vocabulary as a part of the teaching.

We conclude that Searle's error in the design of the Chinese Room is first, allowing direct access to the mechanism while it is part of the Room and, secondly, failing to provide a Chinese-to-English foreign language dictionary. By design this Chinese Room of Searle's speaks English, but, also by design, it cannot give an account, in English, of its Chinese understanding.

THIS PAGE IS LEFT BLANK INTENTIONALLY

REPORT DOCUMENTATION PAGE

DRIC Reference Number (if known) ....

Overall security classification of sheet ....Unclassified...

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the field concerned must be marked to indicate the classification eg (R), (C) or (S).

Originators Reference/Report No. MEMO 4391	Month NOVEMBER	Year 1990

Originators Name and Location

RSRE, St Andrews Road
Malvern, Worcs WR14 3PS

Monitoring Agency Name and Location

Title

ON UNDERSTANDING IN COMMUNICATING SYSTEMS AND THE ERROR IN THE CHINESE ROOM THOUGHT-EXPERIMENT

Report Security Classification UNCLASSIFIED	Title Classification (U, R, C or S) U

Foreign Language Title (in the case of translations)

Conference Details

Agency Reference	Contract Number and Period
Project Number	Other References

Authors WHITE, T A D; GIESS, S C	Pagination and Ref 23

Abstract

John Searle has proposed the Chinese Room thought-experiment to refute the notion of what he calls strong AI (Artificial Intelligence), that is that thinking is merely the manipulation of formal symbols.

We discuss the Chinese Room thought-experiment, and show that Searle's argument is in error.

We show, therefore, that the Chinese Room thought-experiment provides no grounds for the conclusion that computer hardware, solely on the basis of running computer software, can never be regarded as intelligent.

Abstract Classification (U,R,C or S)
U

Descriptors

Distribution Statement (Enter any limitations on the distribution of the document)

UNLIMITED

THIS PAGE IS LEFT BLANK INTENTIONALLY