Is it Enough to Get the Behaviour Right?

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 3A6
hector@cs.toronto.edu

Abstract

This paper deals with the relationship between intelligent behaviour, on the one hand, and the mental qualities needed to produce it, on the other. We consider two well-known opposing positions on this issue: one due to Alan Turing and one due to John Searle (via the Chinese Room). In particular, we argue against Searle, showing that his answer to the so-called System Reply does not work. The argument takes a novel form: we shift the debate to a different and more plausible room where the required conversational behaviour is much easier to characterize and to analyze. Despite being much simpler than the Chinese Room, we show that the behaviour there is still complex enough that it cannot be produced without appropriate mental qualities.

In this paper, we will consider the issue of the relationship between *external behaviours* and *mental qualities*. The external behaviours we have in mind are the linguistic responses in an intelligent conversation. The mental qualities we have in mind are things like knowing how to speak a language, or understanding what is being said, or even being intelligent (none of which we will need to distinguish for now). The fundamental issue we intend to investigate is this:

> *When can we justifiably draw conclusions about mental qualities like these, given external behaviour that is indistinguishable from that of a person?*

In a sense, this question is not really part of AI. One definition of AI is that it is "the study of intelligent behaviour achieved through computational means" [Brachman and Levesque, 2004]. From this point of view, AI research is about getting the behaviour right and nothing more. The question above goes beyond this and asks what conclusions we can draw should we ever get the behaviour right.

The Turing Test [Turing, 1950] and the Chinese Room [Searle, 1980] are two thought experiments designed to help us understand this issue. To recap the positions very briefly, we have Turing who says (roughly) that the mental vocabulary above is too vague and open to interpretation to be worth arguing about. If we are unable to distinguish the responses of an entity from that of a person in an unrestricted conversation (in what Turing calls the Imitation Game), that ought to be enough. In short: if the behaviour in the long run is what it should be, we should be prepared to ascribe the same mental qualities we would to a person. Searle, on the other hand, imagines himself in a room called the Chinese Room where there is a large book. People give him messages written in Chinese, which he does not understand. However, the book in the room tells him what to do with this message, culminating in him writing characters on a piece of paper, the meaning of which he does not understand. He hands the paper back outside the room, and the people there find these responses quite congenial, and in fact indistinguishable from those of a native Chinese speaker. But Searle does not know Chinese. He is producing linguistic behaviour that is indistinguishable from a native speaker's without any understanding. So getting the behaviour right does not justify the ascription of the mental qualities.

So Turing and Searle take opposite positions on the issue above. But one aspect that they both would agree on (one imagines) is this: when we talk about getting the behaviour right, and in a way that is indistinguishable from someone with the appropriate mental qualities, we are not talking about doing so in some *limited* context. All parties agree (or would likely agree) that it is possible to use trickery and other uninteresting means to get the behaviour right in conversations that are restricted enough. For example, a conversant that says nothing but "I love the Yankees!" over and over might be producing conversation that is indistinguishable from that of fanatical baseball fan (possessing mental abilities beyond the four words, one still presumes), but nothing interesting follows from this. Similarly, a test that is limited in advance to a certain number of words may not be enough. What matters to both Searle and Turing and what concerns us here are the conclusions that we would draw about the mental properties of a conversant given that the conversation is natural, cooperative, unrestricted, and as long as necessary.

1 The AI perspective and the Systems Reply

So who is correct here, Turing or Searle? Much of the debate within AI has not really attempted to answer the question.

Regarding the Turing Test, the main discussion has been on whether linguistic behaviour is enough, or whether a more comprehensive notion of behaviour would be a better test [Harnad, 1989]. For example, we might want a notion of behaviour that encompasses broader notions of perception

and action in the world. There has also been discussion on whether passing the Turing Test is a suitable long term goal for AI research [Cohen, 2004]. This is especially germane given the Loebner competition [Shieber, 1994], a restricted version of the test that has attracted considerable publicity. By general consensus, the programs that do well in this competition do not tell us much about intelligence, for the reasons mentioned above. But they do tell us something about fooling people. It appears to be more of a case like ELIZA [Weizenbaum, 1966], where a program using very simple means was able to fool people into believing they were conversing with a psychiatrist. All this simply reflects the fact that it is sometimes possible to simulate linguistic behaviour that has been limited in some way (like that of a Rogerian psychiatrist, or a fanatical baseball fan or, for that matter, a person with autism) by philosophically uninteresting means. None of this reflects directly on the (unrestricted) Turing Test itself.

Regarding the Chinese Room, much of the discussion within AI has been to dismiss it (sometimes quite impatiently) as concentrating on the wrong question: what matters is not whether or not intelligent behaviour is evidence for mental qualities, but rather *how* or even *if* the behaviour can be produced at all, that is, the AI question.

When the former question is addressed, it is typically along the lines of the *Systems Reply* [Searle, 1980]. The argument is that although Searle himself does not know Chinese, the system consisting of Searle together with the book does. For computer scientists, this is a natural notion: Searle is the executor of a program (written in the book), and although the executor does not have a certain ability, it can execute programs that give it that ability. Searle had already anticipated the Systems Reply and had a ready answer: He asks us to extend the thought experiment somewhat, and imagine that he memorizes the contents of the book and then discards it. He still does not understand the Chinese, yet can generate the same linguistic behaviour. But now there is no longer a system consisting of him and the book; there is just him. He is generating behaviour that is indistinguishable from that of a native Chinese speaker without knowing Chinese.

But is Searle's answer to the System Reply really the final word on all this? Is there anything new to say after all this time? For many (especially outside of AI), the debate is over: Searle wins. The Chinese Room even appeared in *Scientific American* as some sort of discovery, like Einstein's thought experiment about travelling near the speed of light.

Rather than throwing in the towel for good, what we will try to do here is to show that Searle's answer to the Systems Reply does not really work: to imagine Searle behaving in a convincingly Chinese way without knowing Chinese involves having to make certain assumptions about the book he has memorized that will be seen to be untenable.

2 Type 1 and Type 2 books

To get started, first observe that whether or not in memorizing the book you end up actually learning Chinese depends on the book. For some books, *no*; but for others, *yes*. Call the books Type 1 and Type 2, respectively.

While Searle clearly wants us to imagine a Type 1 book, here is an example of a Type 2 book: The instructions at the start would say "The marks on the paper you will receive will be a question or statement in Chinese. Use the rest of this book to translate it into English. Then formulate a response, translate it back into Chinese using the rest of this book, print your Chinese response on the paper, and hand it back." The rest of the book would be an elaborate English-Chinese-English manual, with plenty of pictures of Chinese characters, vocabulary, grammar, and examples of usage. A small book might lead to stilted Chinese like that of a first-time tourist, perhaps. But a larger book, with extensive examples of usage, ought to lead to fairly natural Chinese.

Of course, AI supporters get no comfort from a Type 2 book like this one. It suffices to teach Chinese, but it teaches Chinese *as a second language.* It tells us how to answer questions about dogs, say, by connecting the Chinese symbol for dog to the English word "dog," relying on the fact that we *already understand* what the word "dog" means. In a sense, the challenge of AI is to come up with some sort of book for Chinese as a *first* language.

I do not intend to argue from the philosophical armchair about the prospects of AI eventually achieving this goal. The intention here is more modest; I want to argue for this:

There are no Type 1 books for Chinese!

Note that the truth of this claim will still be sufficient to refute Searle's answer to the Systems Reply: if there are no Type 1 books for Chinese, then Searle's assertion that he would not understand Chinese *after learning the book* is just wrong.

But how could we possibly "prove" such a claim? Without knowing what a book for Chinese would need to be like, we are in no position to assess what it would be like to learn it. All we can do is wave our hands. This is maybe why direct philosophical arguments in the past about the Chinese Room have been so stupendously unconvincing: your views about what such a book would have to be like may not coincide with mine. To get beyond these obstacles and see the issues more clearly, we propose moving away from the Chinese Room to a related but simpler thought experiment. We will return to the Chinese Room briefly at the end.

3 The Summation Room

So imagine a Summation Room. Inside the room is a person who does not know how to add numbers. (It might be more realistic to have a person who does not know how to take square roots, since most people do know how to add, but do not know how to find a square root without a calculator. But addition will do.) Messages are passed to the person on a sheet of paper containing a list of twenty numbers, each of which has ten digits. The book inside the room, called Book A, is a very large one: it has ten billion chapters, and each chapter has ten billion sections, and each section has ten billion subsections, and so on up to depth twenty.

The preface of Book A has the following instructions:

> *Take the first number in the list of twenty and go to that chapter; then take the second number in the list and go to that section; then take the third number and go to that subsection, and so on until all twenty numbers have been used up. At the end of this process, there will be a number written in the book with*

at most 12 digits. Write that number on a slip of paper and hand that message back outside the room.

We assume, of course, that unbeknownst to the person in the Summation Room, the book is constructed in such a way that the 12-digit entries in the book are in fact the *sums* of the twenty numbers that led to the entry.

I take it as uncontroversial that the person following the procedure in Book A is not adding. He is producing the correct sums, of course, but only by looking them up. This is no different from phoning a friend and getting the answers from her. And what if the person were to somehow memorize the contents of Book A and follow the instructions in his head? To an external observer, the person is examining the numbers, reflecting for a while, and then writing down their sum. From the outside, it looks just like the numbers are being added. But they are not; it is only a simulation.

So Book A is Type 1 (where *knowing how to add* is substituting here for *knowing how to speak Chinese*) and we can now see Searle's argument very clearly. Paraphrasing Searle, we might say that it is possible to produce behaviour that is indistinguishable from someone who knows how to add, without thereby knowing how to add. Or, more forcefully: Any research program that claims to provide insight into how people are able to add by merely simulating their ability to produce appropriate responses is by itself inadequate.

So it seems at first blush that Searle is right, and Turing wrong: it is possible to simulate the behaviour of someone who knows how to add without having that mental quality.

This, at least, is the argument.

3.1 The problem

The problem with this argument is that it skirts one important consideration: *Book A cannot exist.* As described, Book A would have to contain 10^{200} entries, and our physical universe only has about 10^{100} atoms. So each and every atom in our universe would have to magically hold 10^{100} numbers, that is, another universe worth of numbers! In fact, a storage device with just 10^{20} numbers is at the limit of what we can build today. (For instance, an exabyte = 10^{18} bytes is apparently more than what Google currently uses.) And the ability to store 10^{20} entries would only allow us to handle the answers to the sums of *two* 10-digit numbers! Typical home or office computers would have a hard time storing the answers to the sums of two 6-digit numbers. And books, which are much less dense, would hold less.

So even within its own context as a thought experiment, Book A cannot be real. In the case of the Chinese Room, Searle does not say much about the book itself, steering our attention instead towards much meatier topics like syntax and semantics, meaning and formality, and so on. But this is just misdirection. The problem is quite evident with Book A. We may as well imagine that the room contains a copy of the *Junior Woodchuck Manual*.[1] It is true, and here we fully agree with Searle, that *if* we had Book A, then we could simulate addition without knowing how to add; the problem is that *we cannot have this book* any more than we can have a perpetual motion machine or magical Chinese pixie dust.

It might be argued that we do not really need a book that handles *all* possible 10^{200} inputs. We might try to get by with a fraction (say 10^{20}) of the possible answers. The book would tell us that for any other input the answer is *unknown*. But this will not work. The probability that a list of numbers chosen at random from the 10^{200} possibilities is among the 10^{20} selected ones is essentially zero (*i.e.* to 180 digits).

The misdirection involved with the Chinese Room serves to draw our attention away from any such considerations. The move goes something like this:

> *This is a thought experiment and you should not get too hung up on mundane practical details. There is clearly no problem with Book A in principle. And for a smaller list of smaller numbers, it would be easy to put together a real honest-to-goodness physical book like Book A. You're not arguing that it's the size of the input that really matters here, are you?*

This is similar to the "Aunt Bertha" variant of the Turing Test considered by Ned Block [1981]. He argues that if we can bound the length of the Turing Test in advance, then there is a bound on the number of possible conversations, and so a sufficiently large lookup table can pass the test.[2] What is swept under the rug this time is that "sufficiently large" may be "impossibly large" for all but very tiny bounds. This is made very clear in the Summation Room where a list of twenty 10-digit numbers, which would fit quite comfortably on one single page, is already too much for a lookup table.

4 A Type 2 book for addition

Is the idea of the Summation Room itself the problem here? All this misdirecting talk about "sufficiently large" and "possible in principle" can be so distracting that we may fail to make a crucial observation: *There is a different sort of book that will do the job properly* still without assuming that the person in the room knows any arithmetic.

The preface of Book B has the following instructions:

> *You will be handed a list of twenty 10-digit numbers. Follow procedure* **PROC4***, described below, on these numbers. Do what it says, write the number it says to return on a slip of paper, and hand it back.*

Then the book presents four procedures: **PROC1**, **PROC2**, **PROC3**, and **PROC4**. To be very clear about what works here and why, let us go through these instructions in detail.

First, **PROC1**. Near the start of the book, we have on a page a 10×10 array with rows labelled `0` to `9` and columns labelled `0` to `9`, and whose entries are two digit numbers from `00` to `18`. The procedure **PROC1** is given two digits N and M as input: it

[1] The cartoon character Donald Duck had three nephews who were Junior Woodchucks (similar to Boy Scouts). The running gag was that no matter how preposterously unlikely a predicament the Woodchucks found themselves in, their trusty manual, a slim volume they kept in their backpacks, contained careful step-by-step directions on how to deal with it. To the point: in one episode, they use their manual to find out how to converse with space aliens!

[2] This claim is disputed by Savova and Peshkin [2007] on the grounds that the entries in the table depend on the natural, social, and cultural environment, and so cannot be static and fixed in advance. No matter.

says that what you should do is locate the row for N, locate the column for M in the table and then find the intersection. The answer from **PROC1** will be the two-digit number at the intersection of the row and column. (We, who know how to add, can see that the two digit number in the table is the sum of the digits on the row and column.)

Next, **PROC2**. It takes as input a single-digit number N and a 3-digit number ABC, where the A is not `9`. It says: First use **PROC1** with N and C. Suppose it answers PQ. Use **PROC1** again with P and B. Suppose it answers RS. Finally use **PROC1** with A and R and suppose the answer is UV. The answer to return for **PROC2** is VSQ. (We, who know how to add, can see that the number VSQ is the sum of N and ABC.)

Next, **PROC3**. It takes as input a single 3-digit number PQR whose first digit is `0` or `1`, and a sequence of twenty single-digit numbers, and will return a 3-digit answer. Here's what it says to do: We are going to produce a list of twenty-one 3-digit numbers whose first digit will be `0` or `1`. The first one will be PQR. Then use **PROC2** on the first number in the sequence and the first 3-digit number to get the second 3-digit number. Use **PROC2** again on the second number in the sequence and the second 3-digit number to get the third 3-digit number. Keep on using **PROC2** on a number in the sequence and the corresponding 3-digit number, until all twenty single-digit numbers have been used. The answer from **PROC3** is the final 3-digit number. (Again, we can see that the number returned is the sum of PQR and the twenty single digits.)

Finally, **PROC4**. It takes as input twenty 10-digit numbers and returns a 12-digit answer. Here is what it says to do: Start at the rightmost or 10th digit. Use **PROC3** on `000` and all the 10th digits of the twenty numbers to get a 3-digit number ABC. Write C on a slip of paper. This will be the 12th digit of the final answer. Then use **PROC3** on `0`AB and the 9th digits of the twenty numbers to get a 3-digit number DEF. Write F to the left of C on the slip of paper. This will be the 11th digit of the answer. Then continue with `0`DE and the 8th digits and so on. Finally, use **PROC3** on the leftmost first digits of the twenty numbers, to get XYZ, which we write directly on the slip of paper as the 1st, 2nd and 3rd digits. We have now written the 12-digit number to return. (Again we, who know how to add, can see that the 12-digit number returned is the sum of the twenty 10-digit numbers.)

4.1 Addition by the book

What is the difference between Book A and Book B? They both present algorithms of a sort, although the one in Book A uses much more fixed data than the one in Book B. The main difference, however, is that Book B can actually exist. It's quite small, more like a pamphlet, and easy to memorize.

But more importantly, I claim that the person following the algorithm in Book B is not just looking up answers, but is literally adding the numbers. In other words, a person who memorizes the book and learns **PROC1**, **PROC2**, **PROC3** and **PROC4** actually *learns how to add*. So Book B is Type 2.

The main reason we moved from the Chinese Room to the Summation Room was to get to this point. For most of us, the claim that real addition is taking place with Book B is uncontroversial since, putting aside **PROC1**, we were taught to add using procedures much like **PROC2**, **PROC3**, and **PROC4**.

There is one minor weakness in the argument, however, and that is **PROC1**. Most of us learned *multiplication* by first memorizing a 10×10 table. We don't reason that $8 \times 7 = 56$; we just learn it, the same way we learn that `V` stands for 5 in Roman numerals or that "dog" in French is "chien," and then we go on to learn how to handle multi-digit numbers. But the *addition* of single digits, which we all learned at a very early age, may feel somewhat different. In my case, it feels like left-to-right motion on a number line.[3] But this does not really change the substance of the argument.

Note that we are not claiming that the person who memorizes Book B necessarily *realizes* that he is adding numbers. He may never have heard of addition, and Book B does not require him to relate what he is doing to arithmetic or to counting or even to numbers.[4] But he still knows how to add. In general epistemic terms, we should not expect this type of *de re* knowledge to be closed under introspection.[5]

So when it comes to addition anyway, it may seem that Searle is right and Turing is wrong: Book A shows that it is possible in principle to produce sums without knowing how to add. But this "possible in principle" is vacuous: it relies on a book that cannot exist in reality. With Book B, however, the story is different. A person who memorizes Book B actually learns addition, and not merely a simulation of addition that happens to produce the right external behaviour.

5 Are there Type 1 books for addition?

Are we done? Not yet. The central claim from the Chinese Room is that a certain behaviour can be produced by simulation, without the mental quality. We saw that Book A fails to support this claim, since it cannot exist. But Book B does not *refute* the claim either: there might still be Type 1 books that produce the behaviour in some other more devious way.

So we need to ask: is it physically possible to produce the desired sums by means other than addition? This is a trickier question, though decidedly less tricky than it would be in the Chinese Room. Is the person in the room allowed to phone a friend before giving the final answer? Use a calculator? We would likely want to rule out both of these. On the other hand, we would typically allow passive memory aids, like as much pencil and paper as necessary. And once we allow memory aids, we may as well allow "large" ones. So suppose we put as much as we can into a memory aid, and store the sums of any two 10-digit numbers (all 10^{20} pairs of them). As argued above, this is at the very limit of what we can build today, but it is certainly physically plausible.

So we can imagine a realistic (or somewhat realistic) Book C with the following instructions:

[3] We can alter the **PROC1** procedure to make it more like this. First we construct two identical rulers with twenty even spacings on them with labels `00`, `01`, up to `19`. The new **PROC1** then involves aligning the end of the first ruler at some point on the second, sliding along the first, and getting the corresponding label on the second. We omit the remaining details for space reasons.

[4] For example, we might disguise things using symbols that do not resemble our traditional arabic digits, among other things.

[5] A counterexample: Suppose that Bob speaks French, which happens to be Alice's first language. Then Bob knows how to speak Alice's first language, but may not know that he knows this.

You will be presented with a list of twenty 10-digit numbers. You have a device that will give you back an 11-digit answer for any pair of 10-digit numbers.

Let the number S initially be `0000000000`*. Now repeatedly do the following for each number in the list: use the device with the current number in the list and the current S to get an 11-digit answer. Write the 11th (leftmost) digit of the answer on a scrap piece of paper, and let the next S be the number consisting of the remaining (rightmost) ten digits. At the end, write the final S on a slip of paper. This will be the ten rightmost digits of the 12-digit answer that you will return. To get the remaining two digits, go to the next page for more instructions.*

Does a person who learns this procedure (with details for the carry digits on the next page) learn to add? It is certainly not what we were taught as children, which breaks things down to pairs of digits with a memorized 10×10 table. Here, we break things down to pairs of 10-digit numbers and then look up the memorized answers for them.

Although this is a somewhat unusual procedure, I claim that Book C is Type 2 like Book B. Consider an intermediate case. Book D is just like Book B except that it deals with pairs of digits instead of single digits. It uses a 100×100 table (with rows and columns labelled `00` to `99`) and a modified **PROC1** that does a table lookup for any pair of 2-digit numbers to find a 4-digit answer. The **PROC2**, **PROC3**, and **PROC4** procedures in Book D are analogous to those in Book B, except that they deal with the columns of numbers two at a time.

But this way of handling columns does not really change anything. We can think of it as doing base-100 addition: Instead of viewing the twenty numbers as having ten decimal digits, we interpret them as having five base-100 digits, each of which is written with two characters. It is a different way to add, no doubt, and one that might be just right for individuals with better memories than ours. So working in base-10 with the 10×10 table in Book B is not essential. In fact, we can do addition with even *less* memorized than this table. If we do it in *binary*, for instance, we can get by with a 2×2 table. Book C, then, is simply describing base-10^{10} addition using what amounts to a memorized $10^{10} \times 10^{10}$ table.

So if it is not the size of the table that makes the difference, what does? Why do we say that someone who knows Book A is not really adding, but someone who knows Book B is? In my opinion, we are extending a *courtesy* to the individual with Book B. We do not literally mean that he knows how to add, but only that he knows how to add lists of twenty 10-digit numbers. We extend the courtesy because we can see how easy it is to adapt Book B to deal with *any* list of numbers having *any* number of digits. We focus in the Summation Room on making sure we do the right thing for any of the 10^{200} possible inputs, which are enough to ensure that we cannot get by with just a memorized table. But what we are usually looking for in terms of addition is a *general* competence. Book A cannot be adapted to have this general competence (since it must have a fixed number of pages), but Book B can easily be adapted to handle any list of numbers. The resulting Book B′ might have one more page than Book B. Similarly, Books C′ and D′ would be quite similar to Books C and D.

However, we can also see that we would *not* be inclined to extend the courtesy if all the Summation Room had to do was to add *two* 10-digit numbers, and all the person in the room was doing was using the device in one step to obtain the 11-digit answer. If the person were to somehow memorize all 10^{20} answers we would still not say he knew how to add since the general procedure for handling arbitrary lists of numbers is not even suggested by the very simple 1-step procedure he follows.[6] Just as it is physically possible to produce certain restricted forms of conversation by very simple means (including a lookup table or an ELIZA-like trick), it is physically possible to produce the sums of two 10-digit numbers without knowing how to add. Size matters.

But the question remains: is it also possible to produce the correct answers in the case of twenty 10-digit numbers without wanting to label the resulting procedure "addition"? Addition, as it is normally understood, is the process of coming up with the sums of arbitrary numbers by taking the numbers apart, reducing the problem to some simpler primitive additions whose answers are memorized in advance, and then combining the resulting answers together. So my answer is this: once we accept that we cannot look up the answer (simply because the memory would have to be too large), the only alternative is to operate on the numbers by taking them apart, manipulating them, and putting the answers together piece by piece. If this procedure uses no other source of information, works for any list of numbers (or can be trivially adapted to do so), then I claim we would indeed call it "addition." We might say that the procedure in question was roundabout or clever or even bizarre compared to addition as we normally understand it, but it would still be addition.

In fact, there is an instance of this when it comes to multiplication. The way we were all taught as children of multiplying two n-digit numbers requires on the order of n^2 operations (since every digit from one number is multiplied by every digit from the other). But in 1962, a procedure was discovered that only takes on the order of $n^{\log_2 3}$ operations, a phenomenal improvement [Karatsuba and Ofman, 1962]. The procedure in question is quite far from the usual multiplication procedure, but it is still clearly a case of multiplication.

6 Discussion

As far as the Chinese Room is concerned, Searle does have a point: it is possible to fake *certain kinds* of behaviour without having any of the associated mental qualities. But what holds when that behaviour is extremely *simple* (like adding two 10-digit numbers), need not hold as it becomes more *complex* (like adding twenty 10-digit numbers). The mapping from inputs to outputs for the latter is complex enough that there is no plausible alternative but to process the numeric inputs and perform what amounts to addition. There are no Type 1 books for the Summation Room.

What about the original Chinese Room? As we said, at our current level of understanding, we can only wave our hands about the book there (assuming one could even exist). But

[6] Perhaps a better way of putting it is to say that the person would merely learn the *addition table* for base-10^{10} arithmetic.

while it is true that some forms of conversation are even simpler than adding twenty 10-digit numbers by virtue of being limited enough in length, in scope, or in tone (*e.g.* the fanatical baseball fan above), an unrestricted, long-term conversation in Chinese would surely not fall into this category. The mapping from possible inputs to appropriate outputs in the Chinese Room is so complex compared to the Summation Room that it is ludicrous to imagine that this could be the result of fakery, a trick, a simulation, a lookup table, while the mapping for the Summation Room could not.

This is all we really need. Although analyzing the makeup of the Chinese Room itself is problematic, the Summation Room allows us to do so indirectly. Once we accept that the *Summation Room cannot produce its behaviour without real addition taking place*, we see that the Chinese Room must be similarly constrained. And we stop there. Instead of speculating on what it might or might not be like for Searle to memorize his book, we can shift all the burden back to him:

> *You would have us imagine producing behaviour X using a Type 1 book. But there are no Type 1 books for something as simple as the Summation Room, where the behaviour consists of just a 12-character response to a single 200-character message. Why should we think your behaviour X is easier to fake?*

7 Related work

We are not the first to suggest that there are no Type 1 books for Chinese. In fact, French [2000] argues that there are no books *of any sort* that can do the job Searle asks for, and so concludes (like us) that the thought experiment is vacuous. While there is much to agree with in his paper, I suspect that French is thinking more along the lines of an analogue of the Turing Test, where the job of the book would be to fool an interrogator into thinking she was dealing with a Chinese-speaking *person* (with all the linguistic and non-linguistic life experiences this would normally entail), rather than merely convincing her that the Chinese was being understood.[7]

We are also not the first to use a "complexity" argument to show that lookup tables and the like would need to be too large. French [2000] has a variant dealing with risqué typographical distortions of Chinese characters. Shieber [2007], in his analysis of Block's Aunt Bertha version of the Turing Test, carefully estimates the maximum number of bits that can be physically stored given some very weak assumptions. In all cases, he comes up with estimates that are well below what is needed for the 10^{200} numbers of Book A. Complexity issues like this and others regarding the Turing Test are reviewed and analyzed in [Korukonda, 2003].

8 Summary and Conclusion

To Turing's argument that it is sufficient to exhibit a certain behaviour for a machine to be considered intelligent, Searle responds that, *no*, we can imagine a man in a room with an instruction book who is behaving in a certain way without any understanding of what he is doing. The thought experiment may seem reasonable enough, but in the end, it all depends on the book. The fact that we can imagine a *magical* book should not mislead us; we need to ask what a *real* book would have to be like, and what we would say of the person who learned it. Searle exploits the fact that we do not yet have a clear picture of what a real book for Chinese would have to be like. So we studied a simpler behaviour here, adding numbers, and considered producing that behaviour using a book.

The result: We saw that it was implausible to imagine a person who could produce sums beyond a certain tiny size without knowing addition. We surmised that it was even less plausible to imagine a person who could produce suitable Chinese responses without knowing Chinese.

So in the final analysis, Turing is right: the behaviour is the real issue. Once we accept that simple-minded tricks will not scale up to account for behaviour at the level of complexity of human intelligence, we are left with a very puzzling but *scientific* question: what will? And like Turing, we can let the philosophy go and get on with it.

[7] This is an important distinction and illustrates most clearly, in my opinion, what is *wrong* with the Turing Test.

Acknowledgments

Although the argument here is not rocket science, I did have a lot of trouble expressing it clearly. I thank the referees and early readers Ron de Sousa, Stavros Vassos and Jim Delgrande for their terrific help. Remaining problems? My bad.

References

[Block, 1981] N. Block, Psychologisn and behaviourism. *Philosophical Review* **90**, 5–43, 1981.

[Brachman and Levesque, 2004] R.J. Brachman and H.J. Levesque, *Knowledge Representation and Reasoning*. Morgan Kaufmann, San Mateo, CA, 2004.

[Cohen, 2004] P.R. Cohen, If not the Turing Test, then what? Invited talk of AAAI-04, *xvi*, AAAI Press, 2004.

[French, 2000] R. French, The Chinese Room: Just Say "No!", *Proc. of the 22nd Cog. Sci. Conf.*, 657–662, Philadelphia, 2000.

[Harnad, 1989] S. Harnad, Minds, machines and Searle. *Journal of Theoretical and Experimental AI* **1**, 5–25, 1989.

[Karatsuba and Ofman, 1962] A. Karatsuba and Y. Ofman, Multiplication of multidigit numbers on automata. English version: *Soviet Phys. Ddady.* **7**,7, 595-596, 1963.

[Korukonda, 2003] A. Korukonda, Taking stock of Turing Test: a review, analysis, and appraisal of issues surrounding thinking machines, *Int. J. Hum.-Comput. Stud.*, **58**(2), 240–257, 2003.

[Savova and Peshkin, 2007] V. Savova and L. Peshkin, Is the Turing Test good enough? The fallacy of resource-unbounded intelligence. *Proc. of the IJCAI-07 Conference*, Hyderabad, 2007.

[Searle, 1980] J. Searle, Minds, brains, and programs. *Brain and Behavioral Sciences* **3**, 417-457, 1980.

[Shieber, 1994] S.M. Shieber, Lessons from a restricted Turing Test. *CACM* **37**(6), 70–78, 1994.

[Shieber, 2007] S.M. Shieber, The Turing test as interactive proof. *Noûs* **41**(4), 686–713, 2007.

[Turing, 1950] A. Turing, Computing machinery and intelligence. *Mind* **59**, 433–460, 1950.

[Weizenbaum, 1966] J. Weizenbaum, ELIZA. *CACM* **9**, 36–45, 1966.