UNIB20005 Language and Computation
Week 9: The Chinese Room

Greg Restall
Philosophy Department

Greg Restall

restall@unimelb.edu.au

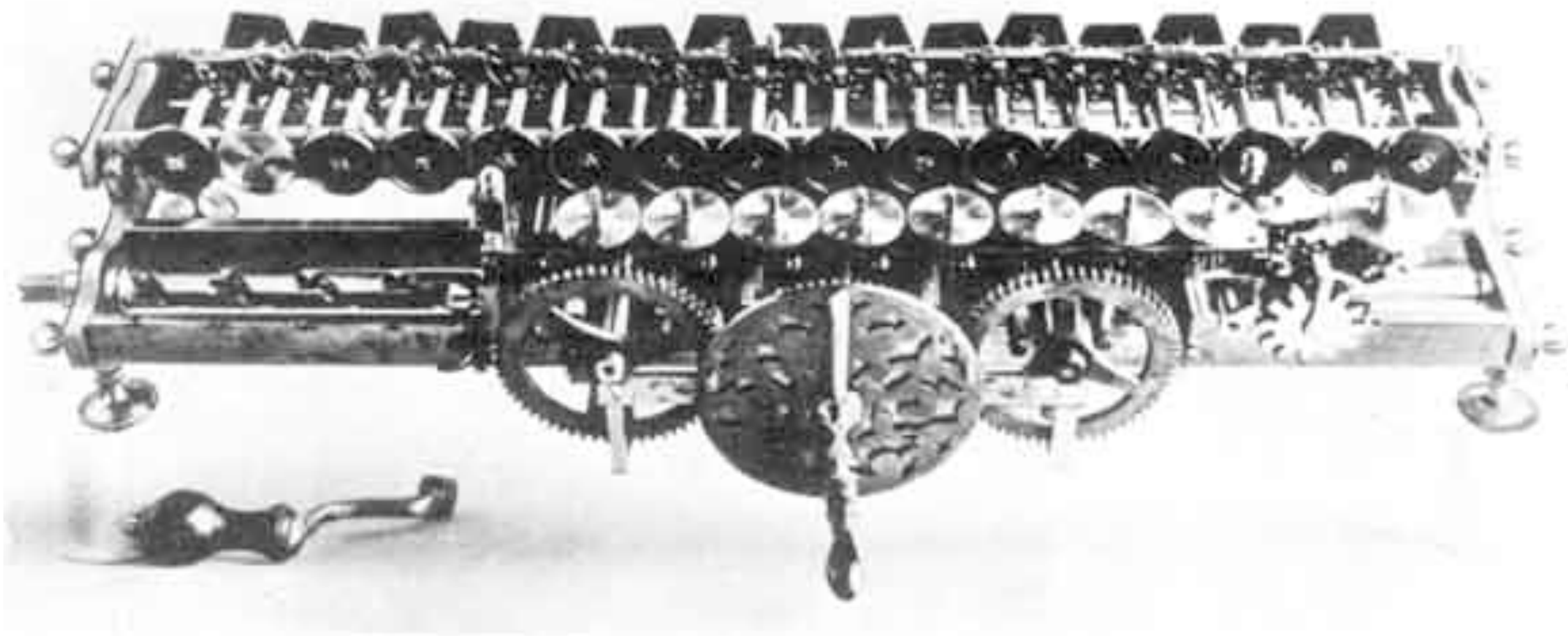
http://consequently.org

Old Quad 203
Mondays and Tuesdays 2–3
and by appointment

The Chinese Room

Gottfried Leibniz

1646 – 1716

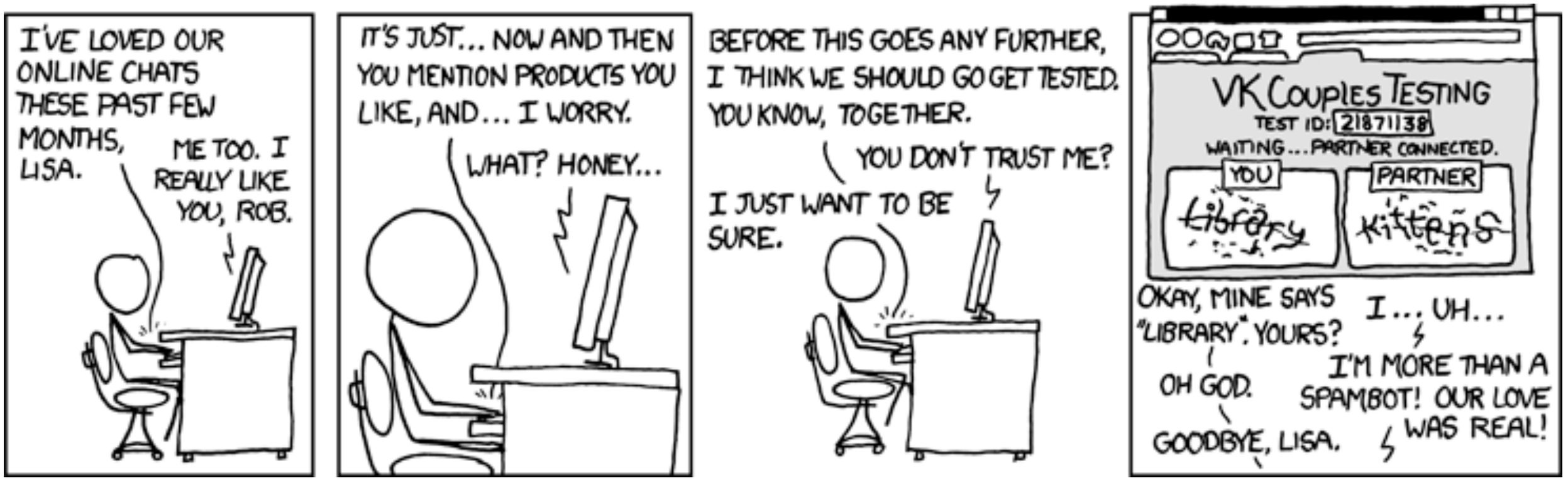


Leibniz’s Mill

Alan Turing

1912 – 1954

Turing Test



http://xkcd.com/632/


hour Toda
Hours
(P)

John Searle

1932 –

THE BEHAVIORAL AND BRAIN SCIENCES (1980) 3, 417–457
Printed in the United States of America

Minds, brains, and programs

John R. Searle
Department of Philosophy, University of California, Berkeley, Calif. 94720

Abstract: This article can be viewed as an attempt to explore the consequences of two propositions. (1) Intentionality in human beings (and animals) is a product of causal features of the brain. I assume this is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality. (2) Instantiating a computer program is never by itself a sufficient condition of intentionality. The main argument of this paper is directed at establishing this claim. The form of the argument is to show how a human agent could instantiate the program and still not have the relevant intentionality. These two propositions have the following consequences: (3) The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program. This is a strict logical consequence of 1 and 2. (4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1. (5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. This follows from 2 and 4.

"Could a machine think?" On the argument advanced here *only* a machine could think, and only very special kinds of machines, namely brains and machines with internal causal powers equivalent to those of brains. And that is why strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking.

Keywords: artificial intelligence; brain; intentionality; mind

What psychological and philosophical significance should we attach to recent efforts at computer simulations of human cognitive capacities? In answering this question, I find it useful to distinguish what I will call "strong" AI from "weak" or "cautious" AI (Artificial Intelligence). According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the

that they can answer questions about the story even though the information that they give was never explicitly stated in the story. Thus, for example, suppose you are given the following story: "A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip." Now, if you are asked "Did the man eat the hamburger?" you will presumably answer, "No, he did not." Similarly, if you are given the following story: "A man went into a restaurant and

http://explodingdog.com/feb26/thisjob.html

The Scenario

Suppose that I'm locked in a room and given a large batch of Chinese writing.

Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that … to me, Chinese writing is just so many meaningless squiggles.

The Scenario

Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch.

The rules are in English, and I understand these rules as well as any other native speaker of English.

The Scenario

They enable me to correlate one set of formal symbols with another set of formal symbols, and all that ‘formal’ means here is that I can identify the symbols entirely by their shapes.

The Scenario

Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch.

The Scenario

Unknown to me, the people who are giving me all of these symbols call the first batch “a script,” they call the second batch a “story.” and they call the third batch “questions.”

Furthermore, they call the symbols I give them back in response to the third batch “answers to the questions.” and the set of rules in English that they gave me, they call “the program.”

The Scenario

Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view — that is, from the point of view of somebody outside the room in which I am locked — my answers to the questions are absolutely indistinguishable from those of native Chinese speakers.

The Scenario

Nobody just looking at my answers can tell that I don’t speak a word of Chinese.

Searle's Question

If there *is* understanding, *where* is it?

Searle's Question

If there *is* understanding, *where* is it?

Since I don't understand Chinese,

following formal rules is not sufficient for **understanding**.

http://explodingdog.com/feb26/thisjob.html

Question 1

What do the programmers need to do?

Question 2

Is there **understanding**, and if so, where?

The Chinese Room Argument (Stanford Encyclopedia of Philosophy)

http://plato.stanford.edu/entries/chinese-room/

STANFORD ENCYCLOPEDIA OF PHILOSOPHY

Cite this entry

Search the SEP

• Advanced Search
• Tools • RSS Feed

Table of Contents
• What's New
• Archives
• Projected Contents

Editorial Information
• About the SEP
• Editorial Board
• How to Cite the SEP
• Special Characters

Support the SEP
• PDFs for SEP Friends
• Make a Donation
• SEPIA for Libraries

Contact the SEP

OPEN ACCESS TO THE ENCYCLOPEDIA HAS BEEN MADE POSSIBLE, IN PART, WITH A FINANCIAL CONTRIBUTION FROM THE UNIVERSITY OF MELBOURNE LIBRARY. WE GRATEFULLY ACKNOWLEDGE THIS SUPPORT.

The Chinese Room Argument

First published Fri Mar 19, 2004; substantive revision Tue Sep 22, 2009

The Chinese Room argument, devised by John Searle, is an argument against the possibility of true artificial intelligence. The argument centers on a thought experiment in which someone who knows only English sits alone in a room following English instructions for manipulating strings of Chinese characters, such that to those outside the room it appears as if someone in the room understands Chinese. The argument is intended to show that while suitably programmed computers may appear to converse in natural language, they are not capable of understanding language, even in principle. Searle argues that the thought experiment underscores the fact that computers merely use syntactic rules to manipulate symbol strings, but have no understanding of meaning or semantics. Searle's argument is a direct challenge to proponents of Artificial Intelligence, and the argument also has broad implications for functionalist and computational theories of meaning and of mind. As a result, there have been many critical replies to the argument.

Stanford Encyclopedia of Philosophy article
by David Cole

Strong AI

Suitably programmed computers can understand natural language and have other mental capacities similar to the humans whose abilities they mimic.

Searle's Argument

1. If Strong AI is true, then there is a program for Chinese such that if any computing system runs that program, that system thereby comes to understand Chinese.

2. I could run a program for Chinese without thereby coming to understand Chinese.

3. Therefore Strong AI is false.

The System Reply

Searle doesn't understand,
but the **system** does.

The Robot Reply

Put the Chinese Room in a body,
and **that** will understand.

Brain Simulator Reply

Something that comprehensively simulates a brain **must** understand.

The Other Minds Reply

We believe **others** understand on no more evidence than in this case.

The Intuition Reply

Why agree with Searle's 'intuition' that he **wouldn't** understand?

There is **no** consensus.

Lessons for us

Syntax is not enough for **semantics**.

Syntax does not determine **reference**.

Syntax does not determine **inference**.

Reference and inference are important parts of **semantics**.

We don’t have a clear idea what **understanding** is.

But the more we can simulate it, **the better**.

We don’t agree now on when we’ll say,

that’s good enough.

Cite this entry

Search the SEP

• Advanced Search
• Tools • RSS Feed

Table of Contents
• What's New
• Archives
• Projected Contents

Editorial Information
• About the
• Editorial
• How to C
• Special C

Support the SEP
• PDFs for SEP Friends
• Make a Donation
• SEPIA for Libraries

Contact the SEP

OPEN ACCESS TO THE ENCYCLOPEDIA HAS BEEN MADE POSSIBLE, IN PART, WITH A FINANCIAL CONTRIBUTION FROM THE UNIVERSITY OF MELBOURNE LIBRARY. WE GRATEFULLY ACKNOWLEDGE THIS SUPPORT.

The Chinese Room Argument

First published Fri Mar 19, 2004; substantive revision Tue Sep 22, 2009

The Chinese Room argument, devised by John Searle, is an argument against the possibility of true artificial intelligence. The argument centers on a thought experiment in which someone who knows only English is alone in a room following English instructions for manipulating Chinese characters, such that to outside observers it appears as if someone in the room understands Chinese. The argument is intended to show that while suitably programmed computers may appear to converse in natural language, they are not capable of understanding language, even in principle. Searle argues that the thought experiment underscores the fact that computers merely use syntactic rules to manipulate symbol strings, but have no understanding of meaning or semantics. Searle's argument is a direct challenge to proponents of Artificial Intelligence, and the argument also has broad implications for functionalist and computational theories of meaning and of mind. As a result, there have been many critical replies to the argument.

Read David Cole's article

Next week:
Dave Ripley
with more on
machine intelligence