foundations of artificial intelligence
fundamental problems

bruno lara

Humboldt-Universität zu Berlin
Institute für Informatik

April 20, 2012

aim

in this section we aim to study the philosophical foundations that have shaped the schools of **artificial inteligence** (?) throughout history

history

turing and the challenge

symbol manipulation

searle thought experiment

history

definitions

definitions

merriam-webster

the ability to learn or understand or to deal with new or trying situations

the skilled use of reason - mental acuteness

the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (as tests)

definitions

oxford dictionary

the ability to acquire and apply knowledge and skills

handbook of human intelligence, cup

goal-directed adaptive behavior

definitions

intelligence journal

a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience

curiosity

ancient desired and thirst for imitating and creating it

ancient - **800 bc**

napata - egypt

statue of *amun* constructed to move its arm and speak to onlookers

ancient - **700 bc**

persia

invented algebra

coined the term *algorithm*

al-Khwarizmi

ancient - **500 ac**

greeks

talos created from a petition from *hephaestus* to *zeus*, to protect *europa*

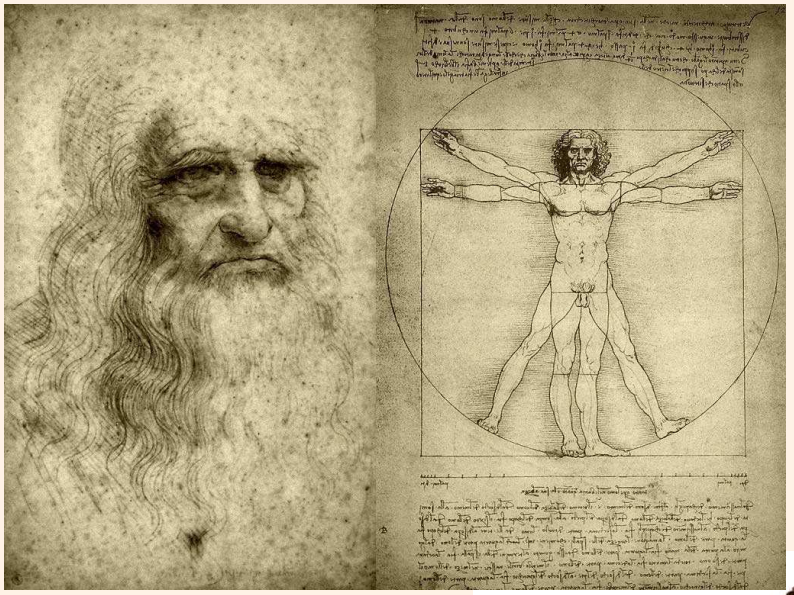
in some versions of the myth, *talos* is forged by the inventor *daedalus*

ancient - **500 bc**

aristotle

syllogistic logic, the first formal deductive reasoning system

ancient - **1515**



ancient - **17th century**

descartes

bodies of animals are nothing more than complex machines

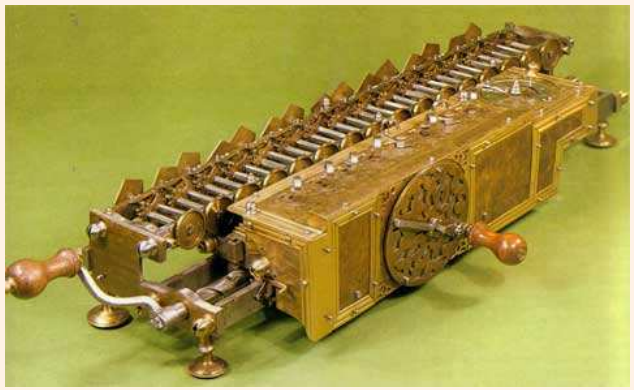
pascal

ancient - **17th century**

thomas hobbes - the leviathan (1651)

mechanistic and combinatorial theory of thinking

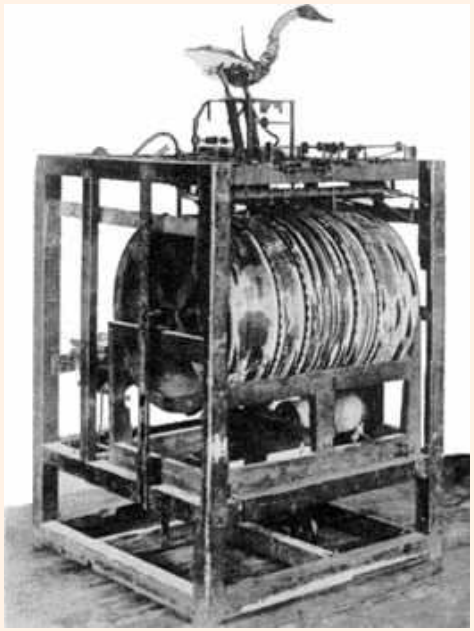
leibniz

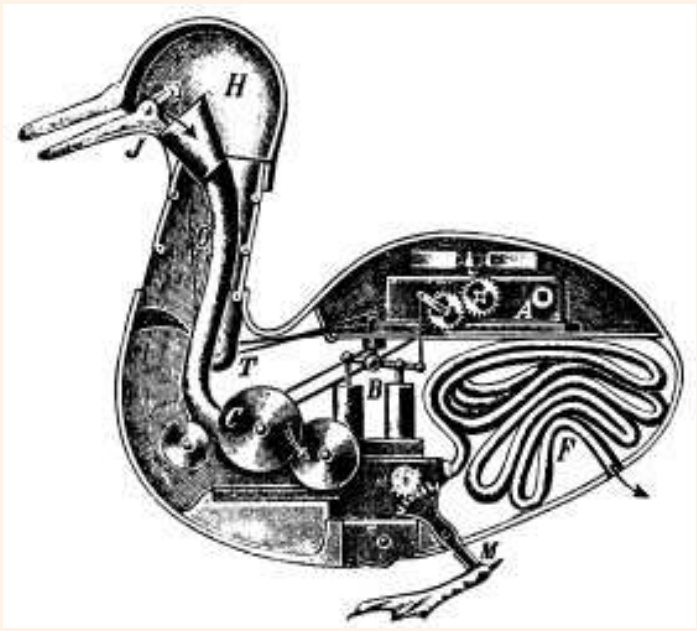


envisioned a universal calculus of reasoning by which arguments could be decided mechanically

ancient - **18th century**

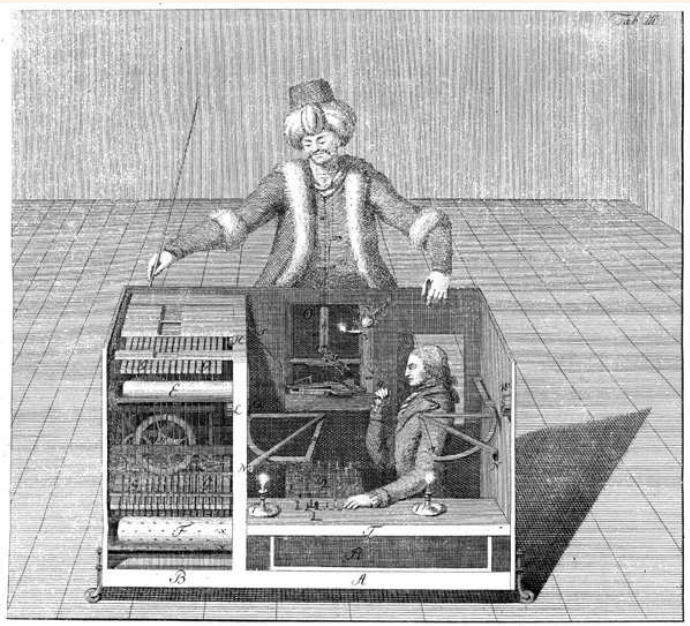
vaucanson





ancient - **18th century**

von kempelen



ancient - **19th century**

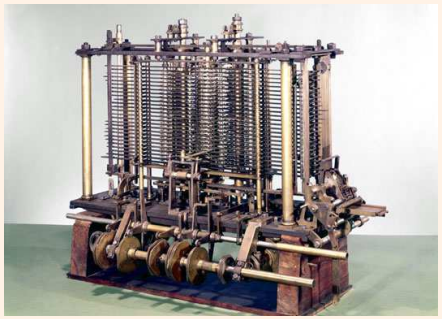
shelley - 1818

frankenstein’s monster

george boole

binary algebra representing “laws of thought”

charles babbage & ada byron



gottlob frege

modern propositional logic - expanded by *russel, tarski, godel, church* and others

modern - **20th century**

principia mathematica - russell and whitehead

revolution for formal logic

karel capek's play

"R.U.R." (rossum's universal robots) - 1921 (published in english, 1923)

mcculloch & pitts - 1943

"a logical calculus of the ideas immanent in nervous activity"

modern - **20th century**

cybernetics - 1943

arturo rosenblueth, norbert wiener & julian bigelow
wiener's book in 1948

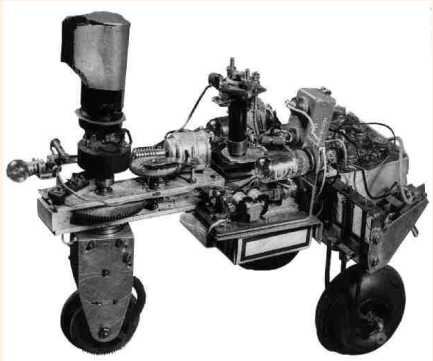
emil post - 1943

proves that production systems are a general computational mechanism

modern - **20th century**

grey walter - 1948 - elsie and elmer

a small number of brain cells
can give rise to complex
behaviors



turing - 1950

“computing machinery and intelligence”

modern - **20th century**

claude shannon - 1950

detailed analysis of chess playing as search
“programming a computer to play chess”

asimov - 1950

three laws of robotics

modern - **20th century**

dartmouth conference - 1956

john mccarthy - comp sci
marvin minsky - cog sci
claude shannon - maths
nathaniel rochester - comp sci
ray solomonoff - comp sci
oliver selfridge - maths
trenchard more - comp sci
arthur samuel - eng
herbert simon - pol
allen newell - maths

main statement

every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it

modern - **20th century**

chomsky - 1957

postulates that language could be analyzed without reference to its content or meaning

syntax independent of semantics

knowledge $\Rightarrow$ represented and analyzed without knowing anything about what was being said

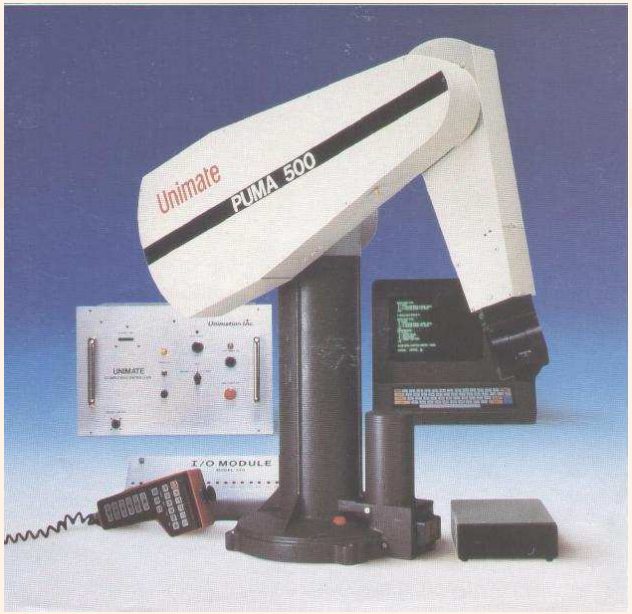
modern - **20th century**

john mccarthy and marvin minsky - 1958

ailab at mit

1962

unimation founded



modern - **20th century**

joseph weizenbaum - 1965

eliza - an interactive program that carries on a dialogue in english on any topic

1980s

neural nets

1990s

deep blue

turing *and the challenge*

turing and the challenge

alan turing

- british mathematician
- consider the *father of comp sci*
- decrypter of *enigma*

turing and the challenge

alan turing

Turing, A. *Computing machinery and intelligence.* Mind. Volume 59. Pp. 433-460. 1950.

turing and the challenge

can machines think?

- define *machine*
- define *think*

turing and the challenge

imitation game

“...expressed in relatively unambiguous words”

- **(a)** a man
- **(b)** a woman
- **(c)** an interrogator

(a)’s aim is to confuse the interrogator **(b)**’s aim is to help the interrogator

(b) >

i am the woman, don’t listen to him!

turing and the challenge

imitation game

- avoid voices, handwritting
- the ideal arrangement is to have a teleprinter communicating between the two rooms

thinking machines

now

- what will happen when a machine takes the part of **(a)** in this game?
- will **(c)** decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?

thinking machines

the machine could

- *intentionally* make mistakes
- falsely take longer times

in short *pretend*

imitation game

to avoid further complications
assume that the best strategy (for the machine) is to try to provide answers that would naturally be given by a man

thinking machines

so, which machines are we talking about?

"..there are already a number of digital computers in working order, and it may be asked, "Why not try the experiment straight away? "

not whether

all digital computers would do well in the game
the computers at present available would do

but whether

there are *imaginable* computers which would do well

thinking machines

digital computer?

machines intended to carry out *any* operations which could be done by a human computer

human computer?

- suppose to be followowing fixed rules
- has no authority to deviate from them
- rules supplied in a book (altered for each job)
- unlimited supply of paper (to perform calculations)

thinking machines

digital computers

- store → paper
- executive unit → carries out the operations
- control → operations are performed in the right order

table of instructions

- book of rules replaced in the machine by a part of the store
- becomes the *table of instructions*

constructing instruction tables ⇒ programming

thinking machines

randomness

digital computer with a random element

- instructions involving the throwing of a die *(or some equivalent electronic process)*
- throw the die and put the resulting number into store 1000
- machines described as having *free will*

"...it is not normally possible to determine **from observing** a machine whether it has a random element ..."

thinking machines

finite machines

- *most* digital computers have a finite store
- *no* theoretical difficulty on unlimited capacity
- only a finite part in use any one time

infinitive capacity computers

thinking machines

babbage - cambridge 1828-1839

- analytical engine
- wheels and cards

use of electricity is of *no* theoretical importance

let’s look for mathematical analogies of function

universality of digital computers

discrete state machines

- sudden jumps or clicks from one quite **definite** state to another
- states are sufficiently different
- negligible confusion of states

universality of digital computers

given the initial state of the machine and the input signals it is always possible to predict all future states

laplace's view

from the complete state of the universe at one moment of time, it should be possible to predict all future states

state of the universe

described by the positions and velocities of all particles

universality of digital computers

discrete state machines

performs predictions nearer to practicability than that considered by *laplace*

the universe as a whole is such that quite small errors in the initial conditions can have an overwhelming effect at a later time

displacement of a single electron by a billionth of a centimetre at one moment might make the difference between a man being killed by an avalanche a year later, or escaping

universality of digital computers

discrete state machines

reasonably accurate knowledge of the state at one moment yields reasonably accurate knowledge any number of steps later

essential property of the mechanical systems

even when we consider the actual physical machines instead of the idealised machines

universality of digital computers

digital computers

fall whithin the class of discrete state machines

the number of states of which such a machine is capable is usually enormously large

universality of digital computers

discrete state machines

possible to predict what it will do

there is no reason why this calculation could not be made by a digital computer

the imitation game could be played with such a machine

universality of digital computers

universal machines

digital computers can mimick any discrete state machine

universality of digital computers

finally

one particular digital computer **c** with

- adequate storage
- suitable speed of action
- an appropriate *programme*

c can be made to play satisfactorily the part of **a** in the imitation game, the part of **b** being taken by a man

universality of digital computers

imitation game

"...expressed in relatively unambiguous words"

- **(a)** a man $\Rightarrow$ a digital computer
- **(b)** a woman $\Rightarrow$ a man
- **(c)** an interrogator

objections

can machines think?

“..i believe to be too meaningless to deserve discussion”

“... i believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted ...”

objections

theological objection

god has given an inmortal soul to every man and woman but not to other animals or to machines

he could give a soul to an elephant if he saw fit

“...such arguments have been found unsatisfactory in the past...”

objections

heads in the sand

the consequences of machines thinking would be too dreadful
let us hope and believe that they cannot do so

we like to believe man to be superior to the rest of creation

“... i do not think that this argument is sufficiently substantial to require refutation...”

objections

mathematical objection

there are a number of results of mathematical logic which can be used to show that there are limitations to the powers of discrete-state machines

from *goedel's*

there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply

the imitation game is a good basis for discussion

objections

consciousness

no mechanism could feel

- pleasure at its successes
- grief when its valves fuse
- be warmed by flattery
- be made miserable by its mistakes
- be charmed by sex
- be angry or depressed when it cannot get what it wants

objections

consciousness

the only way by which one could be sure that machine thinks is to be the machine and to feel oneself thinking

the only way to know that a man thinks is to be that particular man

most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position

objections

disabilities

- be kind, resourceful, beautiful, friendly
- have initiative, a sense of humour,
- tell right from wrong,
- make mistakes, fall in love,
- enjoy strawberries and cream, make some one fall in love with it,
- learn from experience, use words properly,
- be the subject of its own thought,
- have as much diversity of behaviour as a man, do something really new.

objections

disabilities

they are mostly founded on the principle of scientific induction

criticism that a machine cannot have much diversity of behaviour is just a way of saying that it cannot have much storage capacity

disguised forms of the argument from consciousness

objections

lady lovelace

"...the Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform..."

back then they could not imagine a machine could learn

objections

continuity in the ns

a small error in the information about the size of a nervous impulse impinging on a neuron, may make a large difference to the size of the outgoing impulse

which becomes irrelevant for the *imitation game*

objections

informality of behaviour

"...if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine..."

it is *just* a matter of scientifically finding the rules of behaviour

objections

extrasensory perception

- telepathy
- clairvoyance
- precognition
- psychokinesis

overwhelming evidence

learning machines

- initial state of the mind, say at birth
- education to which it has been subjected
- other experience, not to be described as education, to which it has been subjected

similar process to evolution

learning machines

fix a machine with all possible manners of sensing?

restrict it the search in purely intelectual challenges??

intelligence tests

iq

intelligenz-quotient

mean score within an age group is set to 100 and the standard deviation (SD) to 15

approximately 95% of the population scores within two SDs of the mean (iq between 70 and 130)

iq

diferent types of tests

- visual
- verbal
- abstract-reasoning problems
- arithmetic
- spatial imagery
- reading
- vocabulary
- memory
- general knowledge

iq

g general intelligence factor

characterizes the correlations between mental tests, regardless of the tests' contents

iq

stephen jay gould

scientific racism

different types of intelligence

- emotional
- social

symbol manipulation

symbols and search

1975 *acm* turing award lecture

acm association for computing machinery

- allen newell
- herbert simon

Computer Science as Empirical Inquiry: Symbols and Search. Communications of the ACM. March 1976. Volume 19 Number 3.

symbols and search

computer science

the study of the phenomena surrounding computers.

organism to study

not just the hardware, but the programmed, living machine

symbols and search

examples of comp sci as an empirical field

- symbolic system
- heuristic search

symbols and physical symbol ystems

com sci

a fundamental contribution to **knowledge** has been to explain (at a basic level) what symbols are

symbols lie at the root of intelligent action

primary topic of artificial intelligence

symbols and physical symbol ystems

information

processed by computers in the service of ends

intelligence of a system

measured by its ability to achieve ends

laws of qualitative structure

sciences characterize the essential nature of the systems they study

cell doctrine in biology

basic buiding block of living organisms is the cell

laws of qualitative structure

plate tectonics in geology

the surface of the globe is a collection of huge plates

germ theory of disease

most diseases are caused by tile presence and multiplication in the body of tiny single-celled living organisms

atomism

elements are composed of small, uniform particles, differing from one element to another

physical symbol systems

physical

- can be constructed
- it is not restricted to **human** symbol systems

physical symbol systems

a *pss* is composed of two main elements

set of entities

symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure)

processes

- creation
- modification
- reproduction
- destruction

pss

two notions are central

designation

an expression designates an object if, given the expression, the system can either affect the object itself or behave in ways dependent on the object

interpretation

system can interpret an expression if the expression designates a process and if, given the expression, the system can carry out the process

pss

main features

- a symbol may be used to designate any expression whatsoever
- there exist expressions that designate every process of which the machine is capable
- there exist processes for creating any expression and for modifying any expression in arbitrary ways
- expressions are stable; once created they will continue to exist until explicitly modified or deleted
- number of expressions that the system can hold is essentially unbounded

law of qualitative structure

physical symbol system hypothesis

a physical symbol system has the necessary and sufficient means for general intelligent action

necessary

any int sys will prove to be a *pss*

sufficient

any *pss* of sufficient size can be organize to exhibit intelligence

general int action

same scope of intelligence as we see in human action

development of the pssh

pss

instance of a universal machine

the *pssh* implies that intelligence will be realized by a universal computer

development of the pssh

formal logic

progress was first made by walking away from all that seemed relevant to meaning and human symbols

shannon

developed a sys for communication and selection **nothing** to do with meaning

development of the pssh

turing machines and the digital computer

no concept of *symbol* as something that *designates*

inert data is essential to the reduction of computation to physical process

so far, shown that a machine could be run from a description

development of the pssh

store program concept-mid-forties

stored program concept embodies the second half of the interpretation principle: “the system’s own data can be interpreted”

it does not yet contain the notion of designation, of the physical relation that underlies meaning

development of the pssh

list processing-1956

list held addresses which permitted access to other lists

- creation of a genuine dynamic memory structure in a machine
- computer $\longrightarrow$ data types and its operations
- produced a model of designation defining symbol manipulation

evidence

ai

an initial burst of activity aimed at building intelligent programs for a wide variey of almost randomly selected tasks is giving way to more sharply targeted research aimed at understanding the common mechanisms of such systems

evidence

psychology

research focuses on:

- observations and experiments on human behavior in tasks requiring intelligence
- programming of symbol systems to model the observed human behavior

to conclude

pss

study of logic and computers has revealed to us that intelligence resides in physicat symbol systems

to conclude

law of qualitative structure

symbol systems are collections of patterns and processes, the latter being capable of producing, destroying and modifying the former

property of patterns

can designate objects, processes, or other patterns, and that, when they designate processes, they can be interpreted

interpretation

carrying out the designated process

searle's thought experiment

a thought experiment

john searle

- american philosopher
- uc at berkeley
- phil of mind, phil of lang, social phil

a thought experiment

John Searle. *“Minds, Brains and Programs”*, The Behavioral and Brain Sciences.3, pp. 417-424. (1980)

a thought experiment

weak ai

the principal value of the comp in the study of the mind is that it gives us a very powerful tool

strong ai

the appropriately programmed computer really is a mind

the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations

a thought experiment

so...what is searle going to destroy?

- roger schank
- winograd's SHRDLU
- weizenbaum's ELIZA

any *turing* machine simulation of human mental phenomena

a thought experiment

schank's program

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

a thought experiment

schank's program

a man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill

a thought experiment

partisans of *strong* ai claim

the machine is not only simulating a human ability but also

- can literally be said to understand the story and provide the answers to questions
- machine and program explain the human ability to understand the story and answer questions about it

a thought experiment

searle

- i’m locked in a room and given a large batch of chinese writing
- i know no chinese, either written or spoken
- i’m not even confident i could recognize chinese writing as chinese writing distinct from, say, japanese writing or meaningless squiggles
- to me, Chinese writing is just so many meaningless squiggles.

a thought experiment

what else...

- second batch of chinese script
- set of rules for correlating the second batch with the first one (in english)

rules

correlate a set of formal symbols with another set of formal symbols

a thought experiment

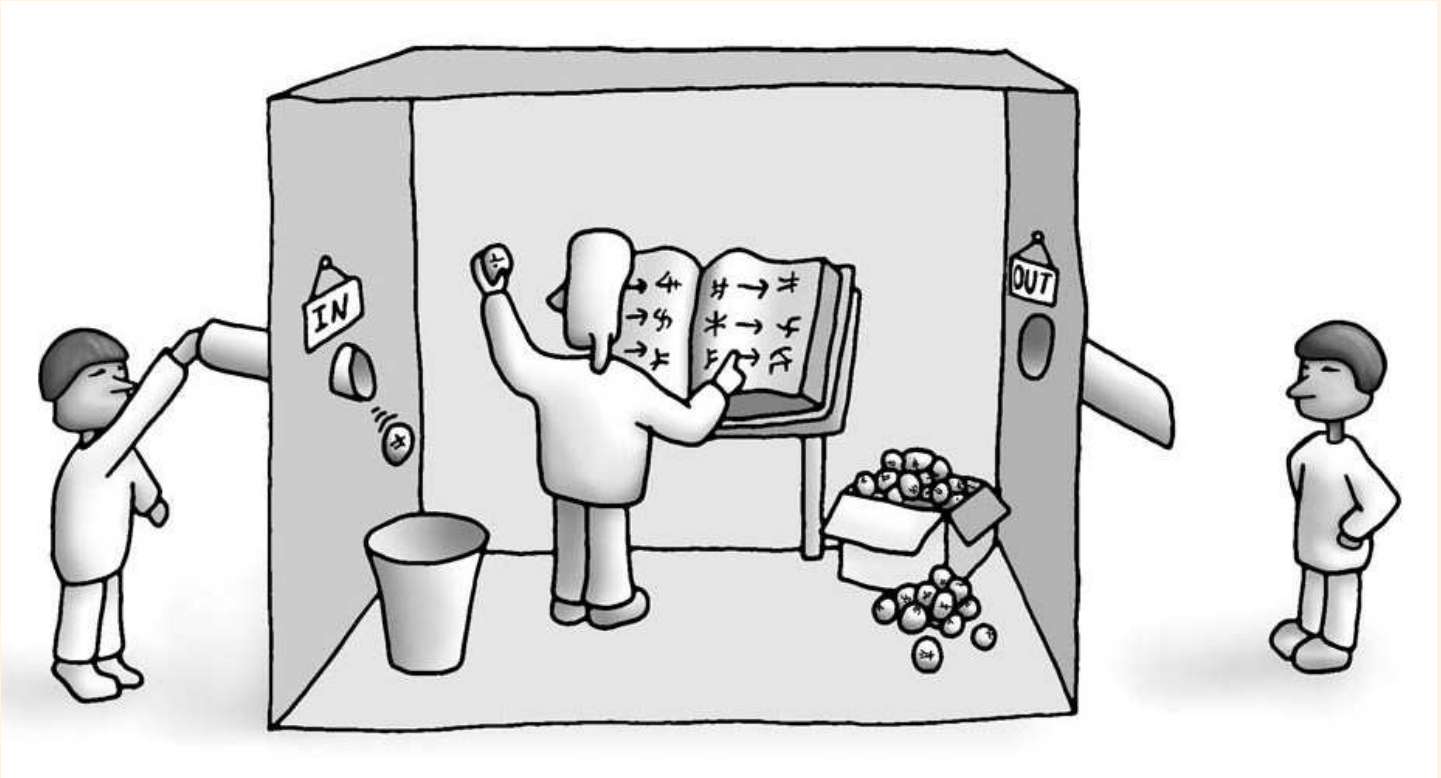
what else...

- third batch of chinese symbols
- set of rules for correlating the thirds batch with the first two (in english)

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

a thought experiment



a thought experiment

- first batch ⇒ script
- second batch ⇒ story
- third batch ⇒ questions
- symbols out ⇒ answers
- rules ⇒ program

a thought experiment

do the same for english and train a lot on both

could someone outside of the room distinguish the difference??

a thought experiment

partisans of *strong* ai claim

the machine is not only simulating a human ability but also

- can literally be said to understand the story and provide the answers to questions
- machine and program explain the human ability to understand the story and answer questions about it

a thought experiment

chinese case

i have everything that ai can put into me by way of a program, and I understand nothing

english case

no reason to suppose that my understanding has anything to do with computer programs, that is, with computational operations on purely formally specified elements

a thought experiment

what is the difference?

a thought experiment

understanding

not the kind meant about a printer, a door, a thermostat

newell and simon

the kind of cognition they claim for computers is exactly the same as for human beings

searle

the programmed computer understands what the car and the adding machine understand

a thought experiment

we extend our intentionality

feature of certain mental states by which they are directed at or about objects and states of affairs in the world

tools are extensions of our purposes

replies

systems reply

understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part

searle's reply

let the person internalize all the elements of the system

replies

- adequacy of the tt
- distinguish mental-nonmental systems
- can not be just on the eye of the beholder

replies

mccarthy →

machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance

searle →

anyone who thinks strong AI has a chance as a theory of the mind ought to ponder the implications of that remark

replies

robot reply

have a robot that has perception, actuators controlled by its computer brain

searle's reply

i am the robot's homunculus, i don't understand anything except the rules for symbol manipulation

replies

brain simulator

simulates the actual sequence of neuron firings at the synapses of the brain of a native chinese speaker when he understands stories in chinese and gives answers to them

replies

searle's reply

as long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states

replies

combination reply

robot + sensors + actuators + brain simulator (of all the synapses)

searle's reply

to a robot whose behavior was indistinguishable over a large range from human behavior, we would attribute intentionality to it

concluding

could a mahine think?

yes, assuming we can produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it

in this case?

could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?

concluding

mind is to brain as program is to hardware

1. only something that has the same causal powers as brains can have intentionality

2. the program is purely formal, but the intentional states are not in that way formal

3. mental states and events are literally a product of the operation of the brain, but the program is not in that way a product of the computer

concluding

which machines could think

only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains

whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality