SYMBOLIC SYSTEMS 100
Introduction to Cognitive Science

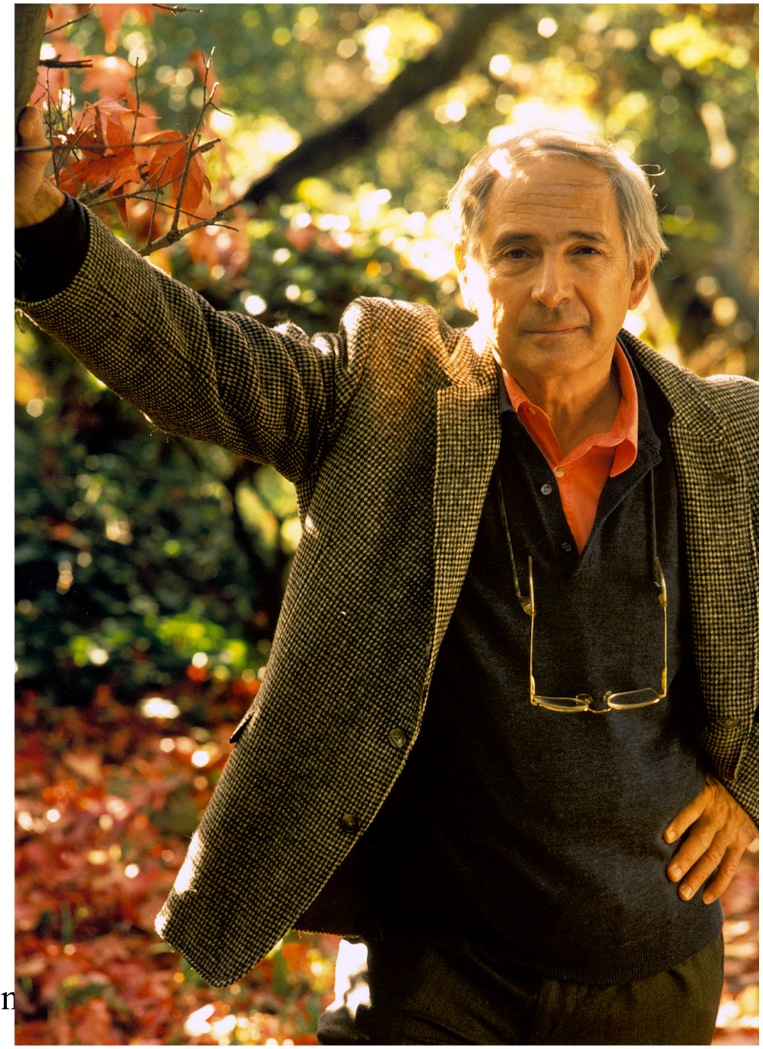
Dan Jurafsky Daniel Richardson

Lecture 2: Searle's Chinese Room and the Turing Machine

IP notice: some slides from David Beaver,
from A. Narayanan, Exeter, and from Polly Huang, EE NTU

The Chinese Room argument

- Searle, John R. 1980. Minds, Brains, and Programs. Behavioral and Brain Sciences.
- Often called the "Chinese Room" paper
- Attacks the claim that "a computer is a mind…" and "can literally be said to understand and have other cognitive states"



Searle wants to argue against 'strong AI'

- Weak AI (maybe we wouldn't call this AI any more): The computer is a useful tool for building a computational model of some cognitive process.
  - We build a model, make predictions, and then test those predictions
- Strong AI (not clear if everyone would call this AI either): "The computer is a mind… Computers given the right programs can be literally said to understand and have other cognitive states."

The historical context

- The early 70's AI work of Roger Schank and students
- "Story understanding"
- Programs that read short stories or news articles and answered questions about them.
  - "A man went into a restaurant and ordered a hamburger. When the hamburger arrived he was very pleased with it; and as he left the restaurant he left a large tip and paid the bill"
  - Q: Did the man eat the hamburger?
  - A: (probably) Yes.

How did these story understanding systems work?

- They were programs using various kinds of knowledge about language and about human behavior
- Such as, for example, a “Restaurant Script”, a representation about what “typically” happens in restaurants (people go in, they order, they eat, they pay)
- This script helped it answer the questions (if you know that someone went in, ordered, got their food, paid, and left, and the story doesn’t say they didn’t eat, you can probably assume they ate)

Searle versus story understanders

- Searle's whole point: to show that just building such a program does not mean the program "understands" in the way that humans do.

Functionalism: A philosophical position

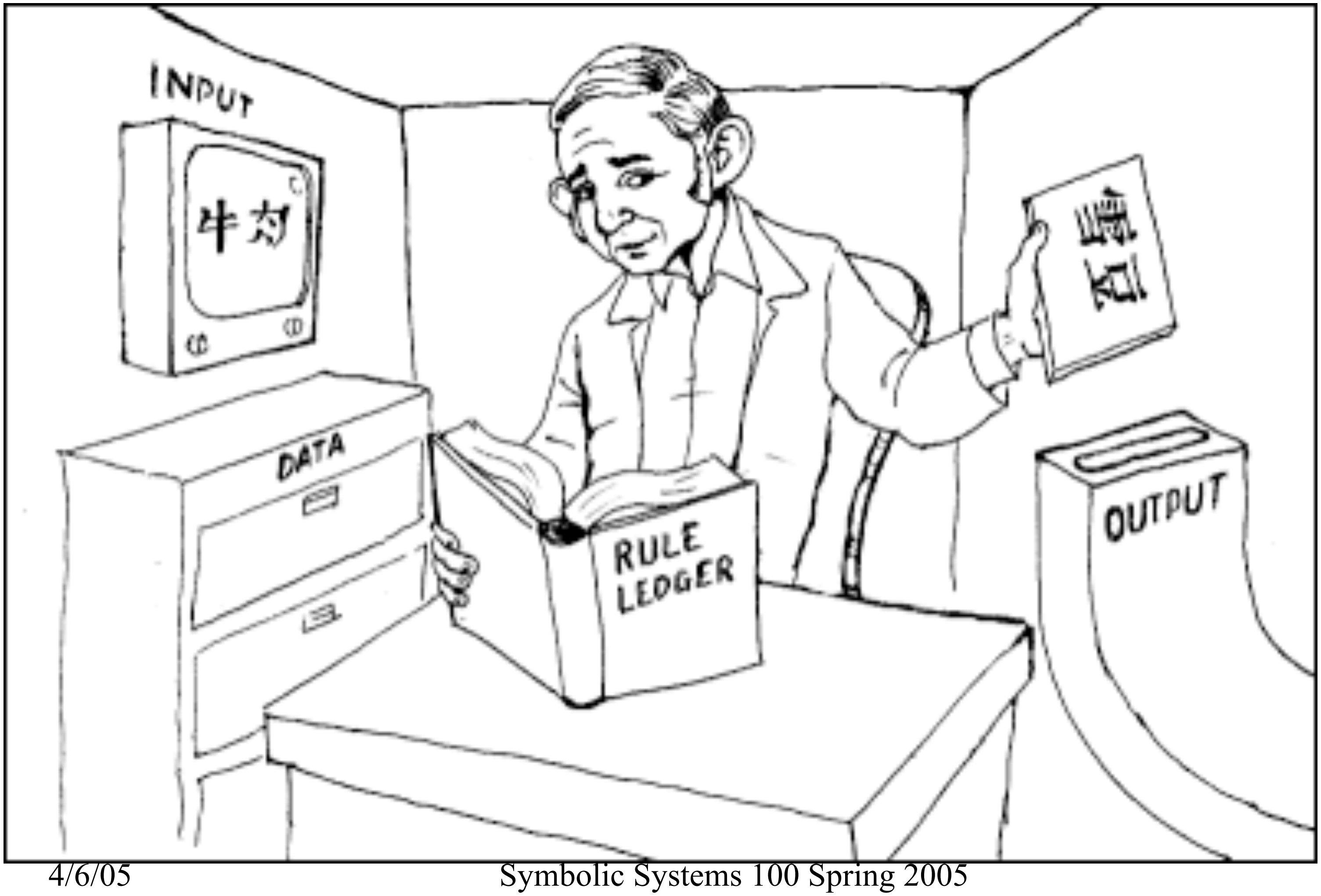
1) mental states are definable independently of the physical (i.e. neural) substrate.
2) They are like “software” and the brain is like “hardware”; you can run the same program on different machines and it’s still functionally identical

- The Strong AI argument that a software AI program could “be a mind and understand” is a functionalist argument

The Chinese Room Experiment

- A Gedankenexperiment (“thought-experiment”)
- Searle, who knows no Chinese, is locked in a room with a batch of Chinese writing
- Now some slips of paper with more Chinese writing are slipped through the door.
- Searle is also given a big rulebook that tells him how to correlate these batches of Chinese symbols with each other and write some new symbols on new slips of paper
- Unbeknownst to him, these are “stories”, and “questions” and he is “answering”.

Searle in the Chinese room



From http://www.unc.edu/~prinz/pictures/

Searle again

From http://www.princeton.edu/~jimpryor/courses/mind/notes/searle.html

“Suppose”, Searle says:

- 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…
- my answers … are … indistinguishable from those of native Chinese speakers.

Compare this, Searle says

- With a more natural situation in English:
- I get stories and questions and I read the stories and answer the questions.
- In the English scenario I am understanding.
- But in the Chinese room, I am not understanding.
- I am just manipulating formal symbols

Searle's point again

- "It seems to me quite obvious in the example that I do not understand a word of the Chinese stories."
- "I have inputs and outputs that are indistinguishable from those of the native Chinese speaker… but I still understand nothing."

A question for you all

- What are the implications of Searle's argument for the Turing test?
- If he is correct, is he saying the Turing test is:
  - Adequate
  - Inadequate
  - Not saying anything about the Turing test

?

Replies to Searle's Argument

- The Systems Reply
- The Robot Reply
- The Brain Simulator Reply
- The Combination Reply
- The Other Minds Reply
- The Many Mansions Reply

The Systems Reply

- "While it is true that the individual person who is locked in the room does not understand the story… he is merely part of a whole system, and the system does understand the story."

Searle's response to the Systems Reply

- "Let the individual internalize all of these elements
- He memorizes the rules in the ledger and the data banks of Chinese symbols
- And he does all the calculation in his head.
- There isn't anything to the system that he does not encompass
- All the same, he understands nothing of the Chinese."

The Robot Reply

- “Suppose we put a computer inside a robot
- This computer would not just take… formal symbols as input and output
- But…the robot does…perceiving, walking, moving about, … eating, … anything you like.
- The robot would… have a television camera… arms and legs…
- Such a robot would have genuine understanding”

Searle's response to the Robot Reply

- "The same thought experiment applies to the robot case…
- Some of the Chinese symbols that come to me come from the television camera,
- And other Chinese symbols that I am giving out serve to make the motors … move the robot's legs…"

Searle's response to the Robot Reply

- Important point: "this reply tacitly concedes that cognition is not solely a matter for formal symbol manipulation, since adds a set of causal relations with the outside world."
- We will return to this idea of "embodiment" on Tuesday (and perhaps you may have thoughts about it today!)

The Brain simulator reply

- “Suppose we design a program that doesn’t represent information about the world, such as the information in Schank’s scripts
- But simulates the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories.”
- Doesn’t this machine understand Chinese?
- At the level of synapses, what… could be different about the program and … the brain?

Searle's response to the Brain Simulator reply

- "Imagine we have the man operate an elaborate set of water pipes with valves connecting them…. Each water connection corresponds to a synapse in the Chinese brain…The man doesn't understand Chinese, and neither do the water pipes."

Searle’s response to the Brain Simulator reply

- Furthermore, says Searle,
- this is a funny response for a functionalist to make!
- The whole point of functionalism, and the “software” metaphor for the mind is supposed to be that “we don’t need need to know how the brain works to know how the mind works”
- Functionalism: there is a level of mental operations consisting of computational processes over formal elements that can be realized in all sorts of different hardwares

The Many Mansions Reply

- Digital computers are just the present state of technology
- “Whatever these causal processes that you say are essential for intentionality, eventually we will be able to build devices that have these causal processes
- And that will be artifical intelligence

Searle's response to Many Mansions reply

- This reply trivializing the project of strong AI
- By redefining it as whatever artificially produces and explains cognition
- The interest of the original AI claim was that it was a precise, well-defined thesis
  - Mental processes are computational processes over formally defined elements
- If that is no longer the thesis, my objections no longer apply because there is no longer a testable hypothesis for them to apply to!

Asking the right question

- ***Could a machine think?***
  - Yes - we are such machines
- ***Could a man-made machine think?***
  - Yes, if we give it appropriate causal powers
- ***Could a digital computer think?***
  - Yes, since even humans can be described as digital computers
- ***Could anything think solely by virtue of being a digital computer?***
  - This is the right question to ask! And the answer, according to Searle, is 'no'

Slide from A. Narayanan

Why not?

- Formal symbol-manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even *symbol* manipulations, since the symbols don't symbolize anything…
- Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them.

Slide form A Narayanan

In other words

- Searle believes that thinking is a physical process
- But the mere manipulation of formal symbols cannot produce genuine thought or intentionality
- Instead, the brain has some sort of special causal power which gives rise to intentionality
- So some sorts of hardware, like human brains, are the kind of thing that can give rise to thought
- Whether thinking takes place depends on exactly the kind of hardware

Furthermore, says Searle:

- The idea that computer simulations of thinking could be the same as thinking ought to have seemed suspicious!
  - No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down
  - Or that computer simulations of a rainstorm will leave us all drenched
- Why should anyone suppose that a computer simulation of understanding understands anything?

Why does anyone believe in strong AI?, asks Searle

- Confusion about ‘information-processing’: we assume that people and computers do it in the same way.
- Residual behaviorism/operationalism - as exemplified by the Turing Test
- Residual dualism: if the mind is a program, it is completely separable from the body

Slide adapted from A. Narayanan

What do you think?

Part II: Turing Machines

Alan Turing again (1912-1954)

- Tuesday we talked about his 1950 paper on the Turing test (and machine learning)
- Today We'll discuss his key theoretical contributions from his 1936 invention of the Turing machine

We keep talking about what "machines" can do

- What do we mean by "machine" or "computer" in the abstract?
- Something like:
  - A mechanism which performs certain kinds of procedures in certain kinds of ways.
- What do we mean by "certain kinds of procedures"?
- We often use the terms "algorithms" or "effective procedures"
- So we really need to ask "What is an algorithm"?

al-Khwarizmi
~780 - 850

What is an "Algorithm" or "Effective Method"?

- **Informal definition:**

  A finite sequence of well-defined instructions for accomplishing some task.

- **Slightly more formally, a method M is called an "effective method" if:**
  - M has a finite number of exact instructions ;
  - M will produce the desired result in a finite number of steps;
  - M can (in practice or in principle) be carried out by a human being unaided by any machinery save paper and pencil;
  - M demands no insight or ingenuity on the part of the human being carrying it out.

al-Khwarizmi
~780 - 850

Algorithms or “Effective methods”

Alonzo Church
1903-1995

Stephen Kleene
1909-1994

And Turing again:

- It turns out there are formalized versions of the informal definition above
- Various formal versions of algorithms proposed in the 1930s by Church, Kleene, Turing, and others (“recursive functions”, “Markov algorithms”, “Post systems”)
- All were proven to be equivalent.
- This led people to accept the Church-Turing Thesis:
    - The informal concept of algorithm is captured by any of the equivalent formalizations.

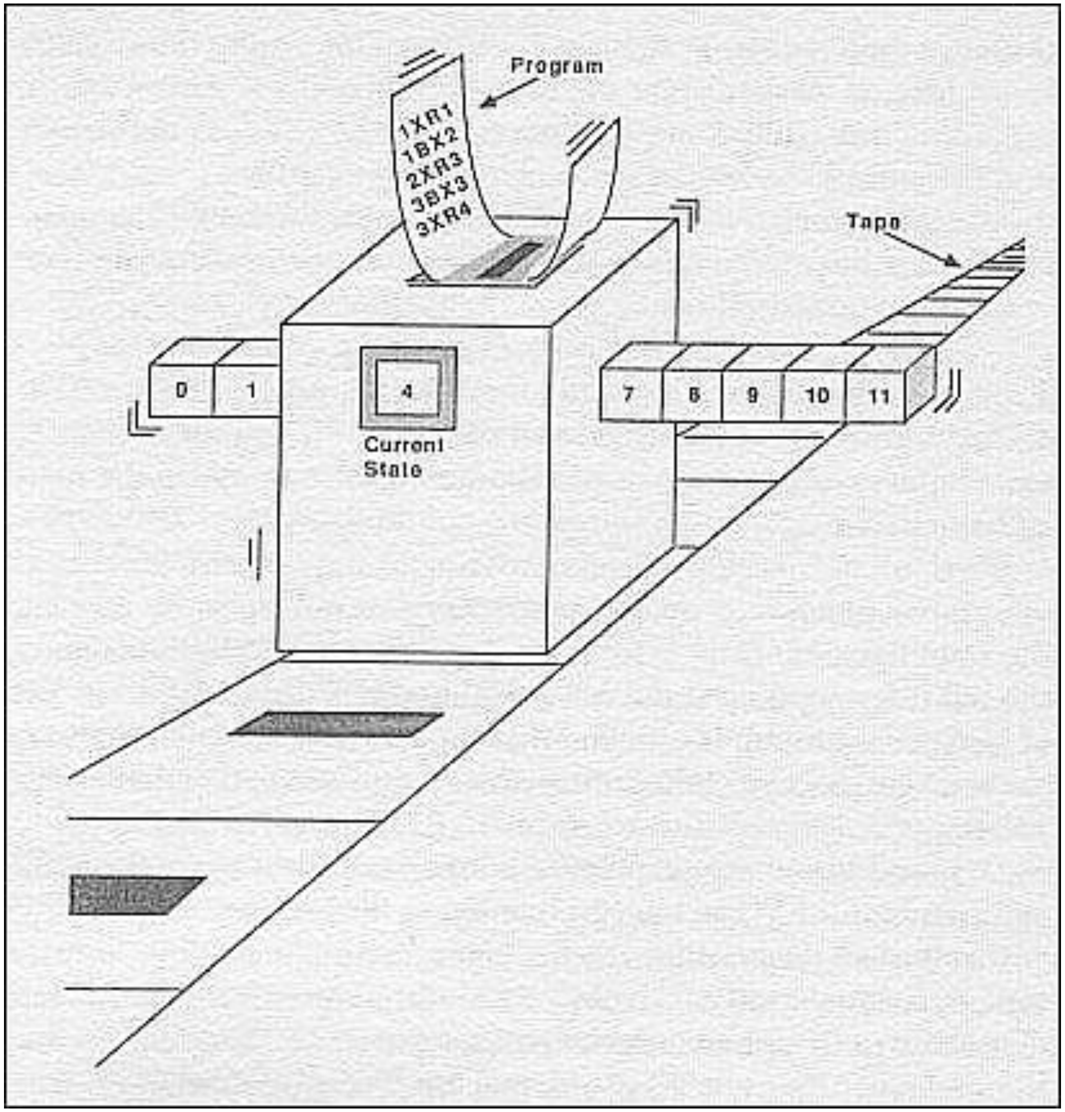
Turing machines

- The informal concept of algorithm is captured by any of the equivalent formalizations.
- The formalization most commonly employed: Turing Machines
- Hence:
  - Turing machines can do anything that can be described as an effective procedure

What is a Turing Machine?

- **An abstract model of a computing machine**
  - **An infinite scannable tape**
    - **with symbols on it**
  - **A moving head**
    - **that reads and writes**
  - **A program telling the head**
    - **which way to move**
    - **and what to write**

A Turing Machine



Turing Machines: more detail

- A read/write head, reading a single cell on an infinite tape with symbols from a finite alphabet;
- A finite number of internal states, one designated as the "start" state;
- Three possible actions:
  - move right,
  - move left,
  - change what's on the tape.
- A finite set of instructions specifying what to do, depending on the state and what is being read.

TMs can be identified with sets of quadruples

- Each quadruple is an instruction that tells the machine what to do depending on what state it's in and what it's reading on the tape.

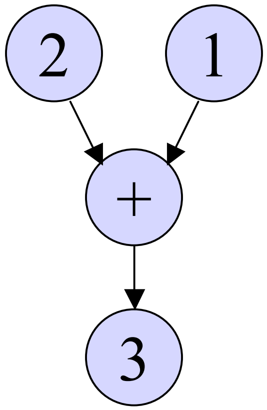
  <STATE, SYMBOL, ACTION, NEW STATE>

  where the possible actions are writing a different symbol in place of the current one, moving left, or moving right.

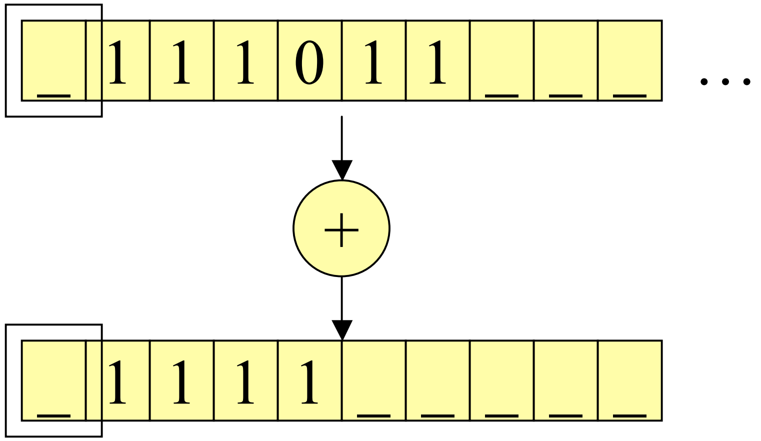
- Each step in a computation involves reading a symbol, carrying out an action, and switching to a new state.

Example: an Adder

2+1=3

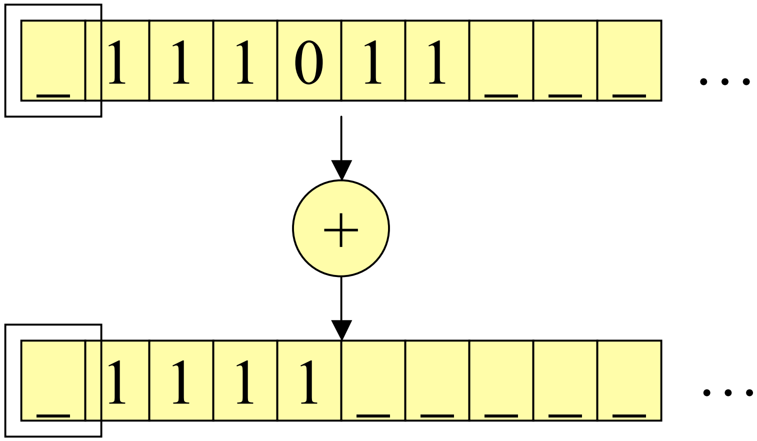


Let n be represented by (n+1) 1’s

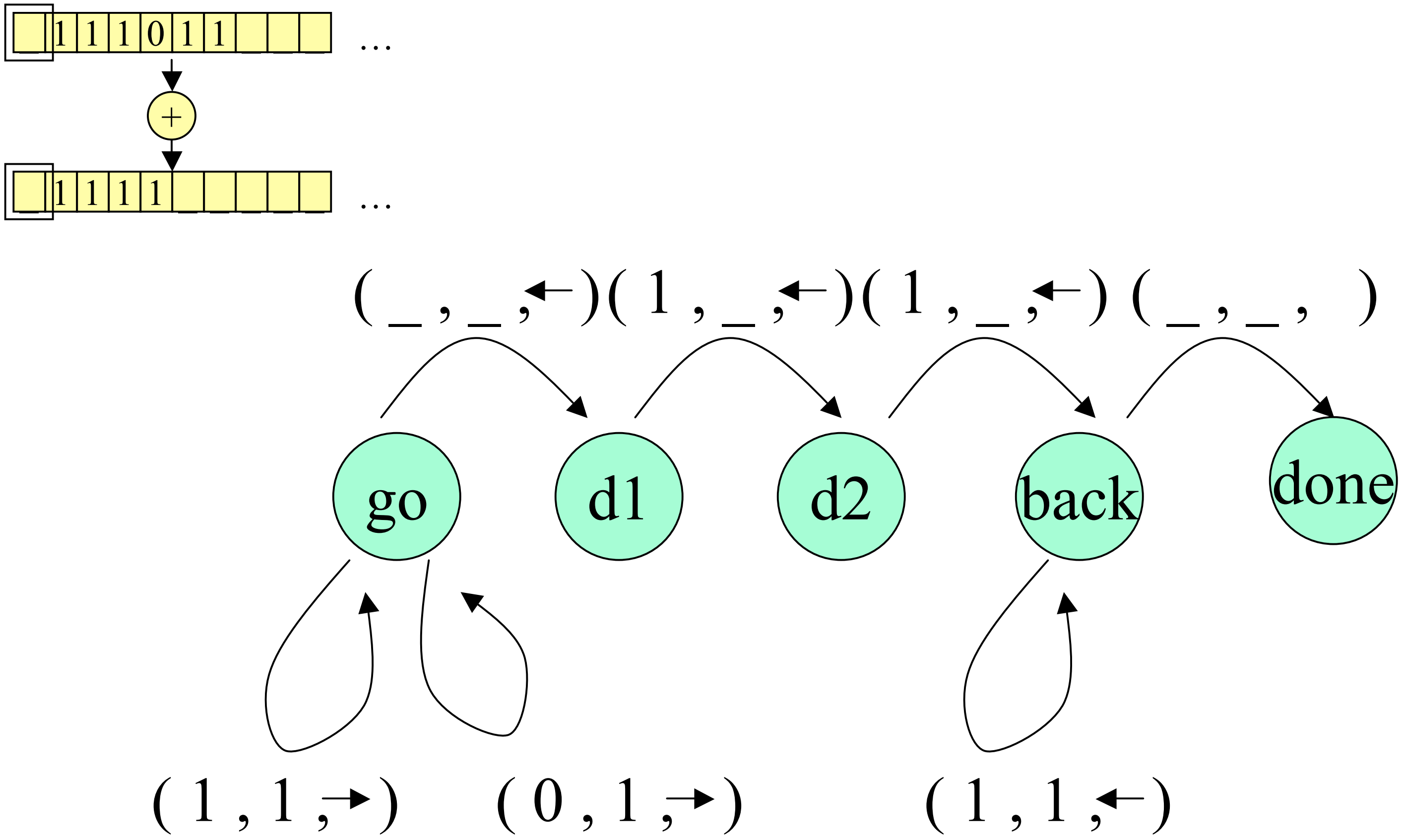


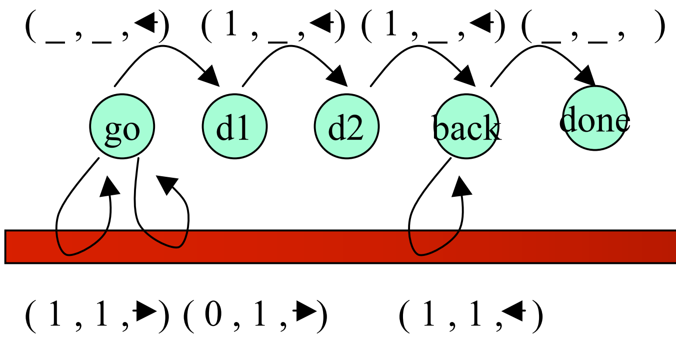
A Turing machine for adding two numbers: 2 steps

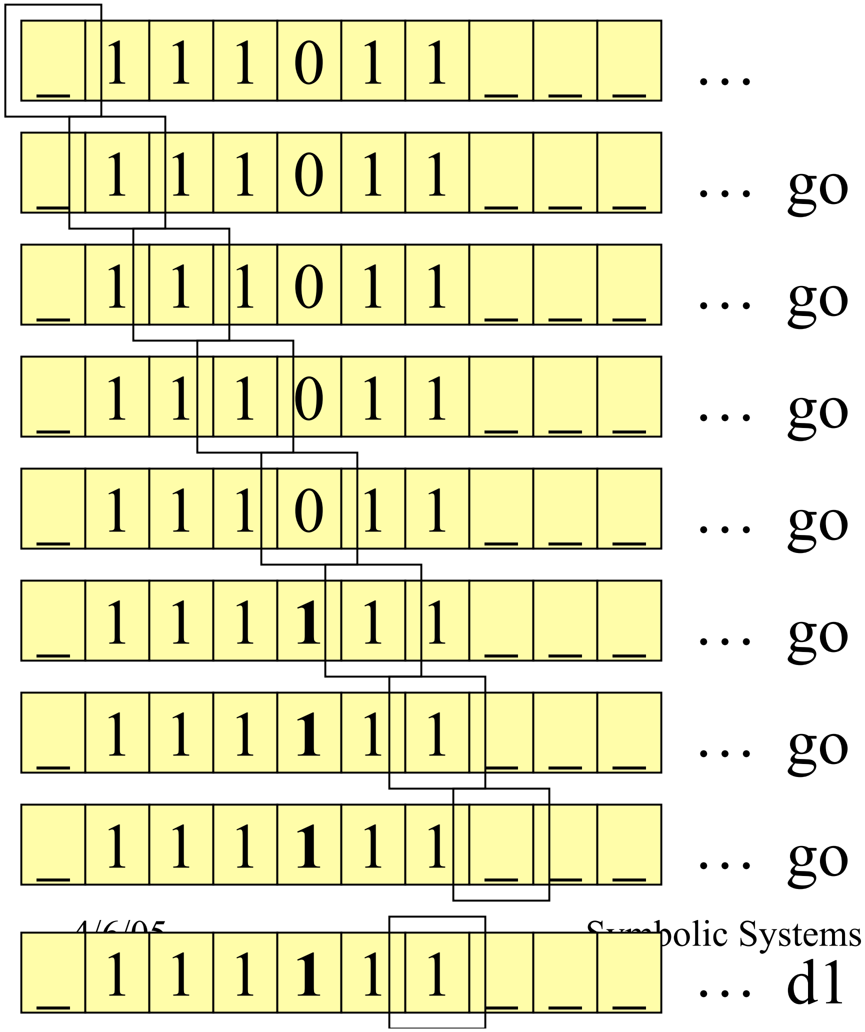
- **Change the 0 to a 1**
- **Erase the two 1's at the end**



State Diagram for the Solution







_ 1 1 1 0 1 1 _ _ _ …

_ 1 1 1 0 1 1 _ _ _ … go

_ 1 1 1 0 1 1 _ _ _ … go

_ 1 1 1 0 1 1 _ _ _ … go

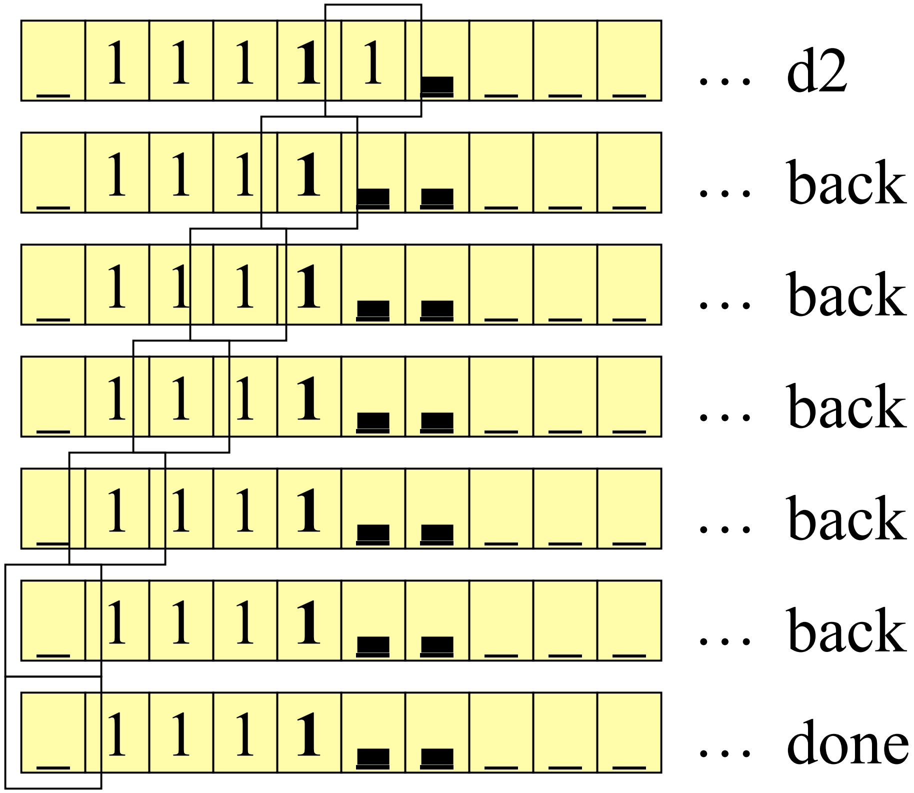
_ 1 1 1 0 1 1 _ _ _ … go

_ 1 1 1 1 1 1 _ _ _ … go

_ 1 1 1 1 1 1 _ _ _ … go

_ 1 1 1 1 1 1 _ _ _ … go

_ 1 1 1 1 1 1 _ _ _ … d1



_ 1 1 1 1 1 ▬ _ _ _ … d2

_ 1 1 1 1 ▬ ▬ _ _ _ … back

_ 1 1 1 1 ▬ ▬ _ _ _ … back

_ 1 1 1 1 ▬ ▬ _ _ _ … back

_ 1 1 1 1 ▬ ▬ _ _ _ … back

_ 1 1 1 1 ▬ ▬ _ _ _ … back

_ 1 1 1 1 ▬ ▬ _ _ _ … done

Wow! 2+1=3

So what?

- Ok, now we have a machine that can add two numbers, now what?
- As we said earlier, we can build a Turing machine to implement any single algorithm
- Furthermore, we can put on the tape not just the data but THE ALGORITHM TOO!
- This means we can build a turing machine to read other turing machines!!!!

The Universal Turing Machine

- Can read a tape that has
  - A representation of a turing-machine program
  - And some data
- And will run the program on the data on the tape!
- So a Universal Turing machine has exactly the power of any complete general-purpose computer:
  - It can run any program we care to write.

The Halting Problem

- Some problems are uncomputable
- For example, it would be nice to be able to look at a Turing machine and an input tape, and just decide if the machine will ever halt, I.e. come to a final state.
- Turing proved that this function is uncomputable
- That is, there is no Turing machine which can tell if any possible Turing machine will halt on a particular input.
- This is an example of a limitation to computation that we talked about last time

Summary

- Strong AI, a subtype of Functionalism, says that
  - The mind can be simulated by a symbolic system
  - Furthermore, this symbolic system itself is also a mind.
- Searle argues that this is just wrong.
- What is this symbolic system that so annoys Searle?
  - A Turing machine
  - A Turing machine can compute anything that a digital computer can
  - Many cognitive scientists, contra Searle, believe that the mind is a symbolic system of this sort.
  - Others believe that a symbolic system is a useful metaphor and conceptual tool