

Limits of symbolic models

LIN 313 Language and Computers
UT Austin

September 5, 2025
Instructor: Gabriella Chronis

Admin

- Homework 1 Due Friday, September 12
- Start working on it early
- come to office hours!

Today

- return concept mapping exercise
- Big questions of Unit 1
 - Why is language a "hard" problem for computers?
 - What makes language unique?
 - How is computation like language?
- today's lecture
 - approach and challenges of symbolic (vs sub-symbolic) AI
 - SHRDLU
 - garden path effects (ambiguity and constituency)
 - winograd schemas (ambiguity)
 - basic difference symbolic and subsymbolic AI
 - deduction and induction
 - rules vs learning

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

John McCarthy, Marvin Minsky, Claude Shannon , Allen Newell, and Herbert Simon (1956)

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that 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. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1 Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2. How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture . From this point of view forming a generalization

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

John McCarthy, Marvin Minsky, Claude Shannon (1956)

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that 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. An attempt will be

problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1 Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2. How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view forming a generalization

Two competing visions of intelligence

Two competing visions of intelligence

From the very beginning, there was a split among AI researchers

- symbolic
 - logic
 - deduction
- subsymbolic
 - induction
 - learning from similar examples

The Dartmouth guys were mostly on the side of symbolic systems

Symbolic AI or 'Good old fashioned AI' (GOFAI)

insight: **formal logical systems** have many of these important properties of language

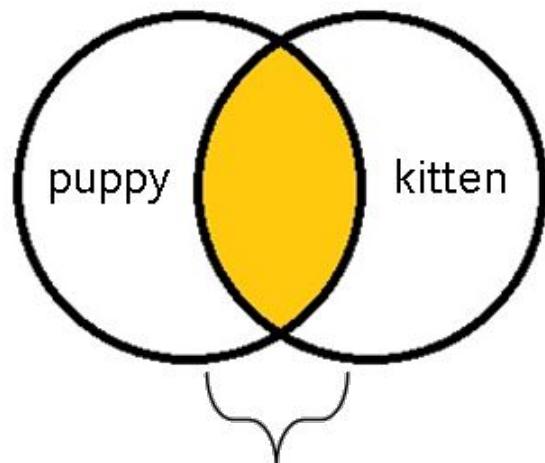
innovation: use computers to model thought using formal languages

- the world consists of symbols (entities, events, properties)
- and predefined relationships between those symbols (subset/superset relationships, causality, etc.)
- programs are rule-based

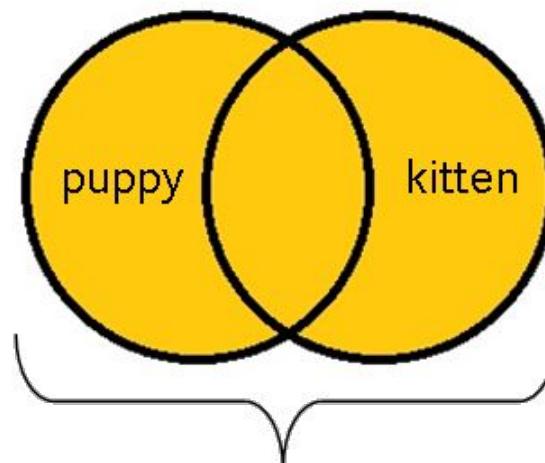
inspiration: **mathematical logic** used to describe mathematical systems

Boolean Logic

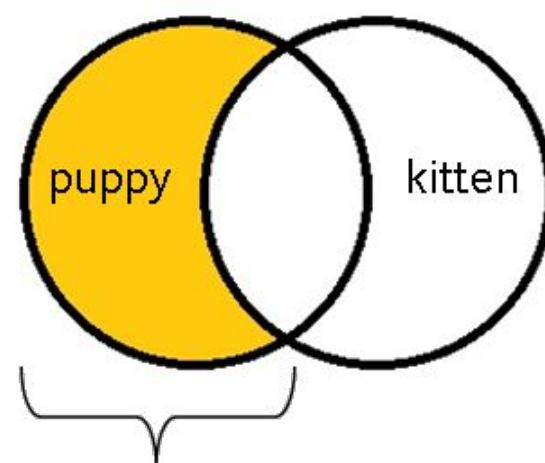
AND



OR



NOT



Aside: how do computers represent text?

Aside: how do computers represent text?

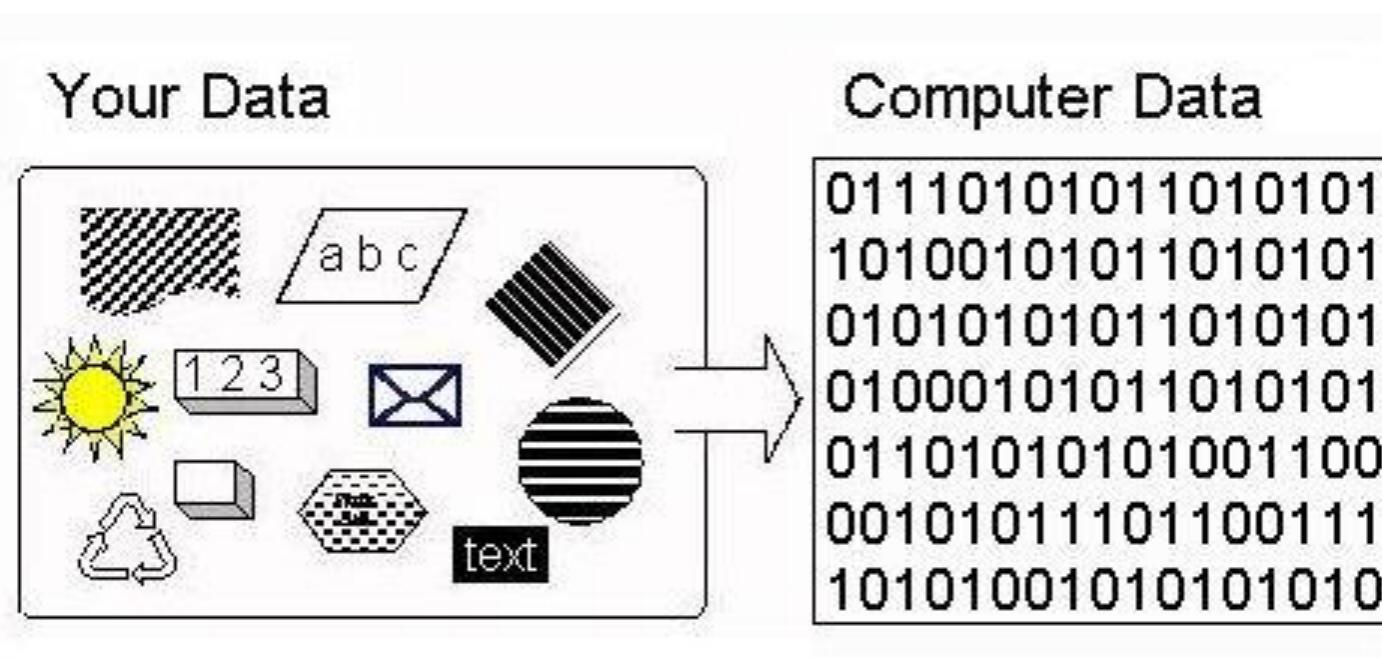
- Let's simplify the problem:
 - How do computers represent any kind of information?

"The files are in the computer."

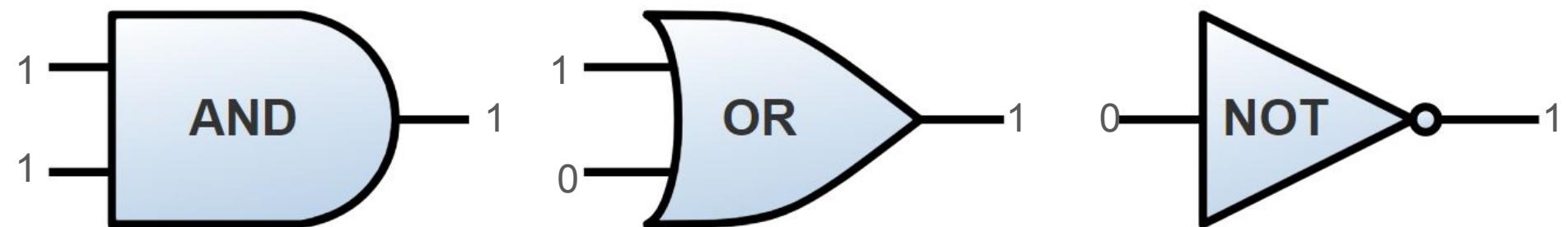
- What does this mean?
 - *What* is in the computer?

Aside: how do computers represent text?

- Let's simplify the problem:
 - How do computers represent any kind of information?



Logic Gates: the building blocks of computers



The General Problem Solver

Simon and Newell (1957, RAND Corporation)

Hypotheses:

- intelligence can be abstracted away from particular problem domains
- thinking is a search through a space of possibilities
 - the solution is the shortest path to the goal
- thinking is essentially logical reasoning about arbitrary symbols

The General Problem Solver

Simon and Newell (1957, RAND Corporation)



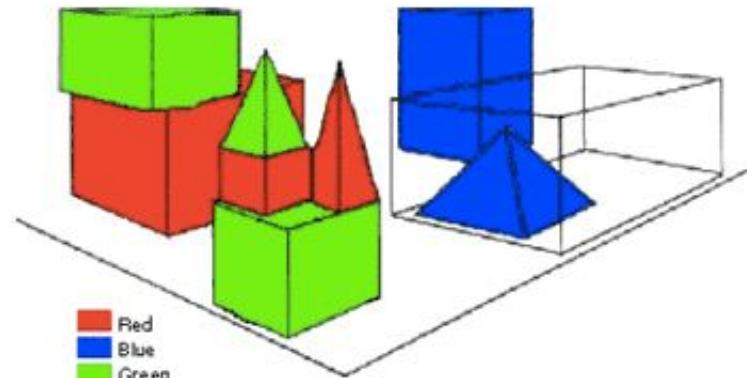
Terry Winograd & SHRDLU

COGNITIVE PSYCHOLOGY 3, 1–191 (1972)

Understanding Natural

TERRY WINOGRAD

*Massachusetts Institute of
Technology
Cambridge, Massachusetts*



Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

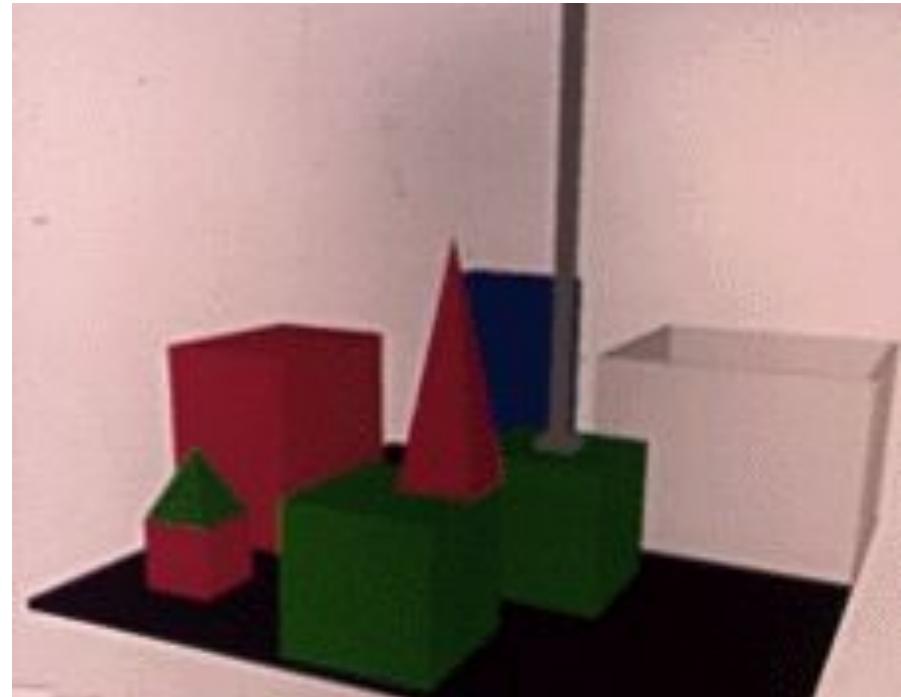
Computer: I don't understand which pyramid you mean.

This paper describes a computer system for understanding English. The system answers questions, executes commands, and accepts information in an interactive English dialog.

It is based on the belief that in modeling language understanding, we

SHRDLU (Terry Winograd, 1968)

- Dialog system for interacting with a “blocks world”
- Very limited domain; easy to encode current state
 - Red, blue, and green things
 - Pyramids and cubes
 - Relative position of objects
- highly complex and sophisticated system
- independent modules for syntactic **parsing**, semantic **interpretation**, reasoning, and planning





Ambiguity of meaning

Person: HOW MANY THINGS ARE ON TOP OF GREEN CUBES?

Computer: I'M NOT SURE WHAT YOU MEAN BY "ON TOP OF" IN THE PHRASE "ON TOP OF GREEN CUBES".

DO YOU MEAN:

- 1 - DIRECTLY ON THE SURFACE
- 2 - ANYWHERE ON TOP OF ?

Elided response

Person: 2

Computer: THREE OF THEM.

Person: HAD YOU TOUCHED ANY PYRAMID BEFORE YOU PUT THE GREEN ONE ON THE LITTLE CUBE?

Computer: YES, THE GREEN ONE.

Person: WHEN DID YOU PICK IT UP?

Computer: WHILE I WAS STACKING UP THE RED CUBE, A LARGE RED BLOCK AND A LARGE GREEN CUBE.

Person: WHY?

Computer: TO GET RID OF IT.

Person: WHY DID YOU DO THAT?

Computer: TO CLEAN OFF THE RED CUBE.

Person: WHY DID YOU CLEAR OFF THAT CUBE?

Computer: TO PUT IT ON A LARGE GREEN CUBE.

Person: WHY DID YOU DO THAT?

Computer: TO STACK UP THE RED CUBE , A LARGE RED BLOCK AND A LARGE GREEN CUBE.

Person: WHY DID YOU DO THAT?

Computer: BECAUSE YOU ASKED ME TO.

Person: HOW DID YOU DO IT?

Computer: BY PUTTING A LARGE RED BLOCK ON THE TABLE ; THEN LETTING GO OF IT ; THEN PUTTING A LARGE GREEN CUBE ON IT; THEN LETTING GO OF THAT CUBE ; THEN PUTTING THE RED CUBE ON THAT CUBE; THEN LETTING GO OF THAT CUBE.

Temporal understanding

Explanation: why

Explanation: how

SHRDLU: example (cont'd)



Duration of events

Person: HOW MANY OBJECTS DID YOU TOUCH WHILE YOU WERE DOING IT?

Computer: FOUR OF THEM.

Person: WHAT DID THE RED CUBE SUPPORT BEFORE YOU STARTED TO CLEAN IT OFF?

Computer: THE GREEN PYRAMID.

Person: THERE WERE FIVE BLOCKS TO THE LEFT OF THE BOX THEN.

Computer: NO, ONLY FOUR OF THEM: THE RED CUBE, TWO LARGE GREEN CUBES AND A LARGE RED BLOCK.

Person: PUT THE BLUE PYRAMID ON THE BLOCK IN THE BOX.

Computer: OK. (does it)

Person: IS THERE ANYTHING WHICH IS BIGGER THAN EVERY PYRAMID BUT IS NOT AS WIDE AS THE THING THAT SUPPORTS IT?

Computer: YES, THE BLUE BLOCK.

Person: DOES A STEEPLE

Computer interrupting: SORRY, I DON'T KNOW THE WORD "STEEPLE".

Person: A "STEEPLE" IS A STACK WHICH CONTAINS TWO GREEN CUBES AND A PYRAMID.

Computer: I UNDERSTAND.

Person: ARE THERE ANY STEEPLES NOW?

Computer: NO.

Person: BUILD ONE.

Computer: OK. (does it)

Person: CALL THE BIGGEST BLOCK "SUPERBLOCK"

Computer: OK.

Person: HAVE YOU PICKED SUPERBLOCK UP SINCE WE BEGAN?

Computer: YES.

Person: WHY DID YOU DROP IT?

Computer: BY "IT", I ASSUME YOU MEAN THE SUPERBLOCK.

I CAN'T EXPLAIN A NON-EXISTENT EVENT.

Person: IS THERE ANYTHING TO THE RIGHT OF THE RED PYRAMID?

Computer: YES, FOUR THINGS: THE BOX, THE RED CUBE, THE BLUE PYRAMID AND THE BLOCK NAMED SUPERBLOCK

Person: THANK YOU.

Computer: YOU'RE WELCOME!

Recognize false assertion

Comparison and quantification

Word and concept learning

Use of learned concept

Naming of objects

Presupposition failure

SHRDLU: example conversation



Person: PICK UP A BIG RED BLOCK

Computer: OK. (does it)

Person: GRASP THE PYRAMID.

Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.

Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Computer: OK. (does it)

Person: WHAT DOES THE BOX CONTAIN?

Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.

Person: WHAT IS THE PYRAMID SUPPORTED BY?

Computer: THE BOX.

Person: HOW MANY BLOCKS ARE NOT IN THE BOX?

Computer: FOUR OF THEM.

Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?

Computer: YES, THE RED CUBE.

Person: IS IT SUPPORTED?

Computer: YES, BY THE TABLE.

Person: CAN THE TABLE PICK UP BLOCKS?

Computer: NO.

Person: CAN A PYRAMID BE SUPPORTED BY A BLOCK?

Computer: YES.

Person: CAN A PYRAMID SUPPORT A PYRAMID?

Computer: I DON'T KNOW.

Person: STACK UP TWO PYRAMIDS.

Computer: I CAN'T.

Person: WILL YOU PLEASE STACK UP BOTH OF THE RED BLOCKS AND EITHER A GREEN CUBE OR A PYRAMID?

Computer: OK. (does it)

Person: WHICH CUBE IS SITTING ON THE TABLE?

Computer: THE LARGE GREEN ONE WHICH SUPPORTS THE RED PYRAMID.

Person: IS THERE A LARGE BLOCK BEHIND A PYRAMID?

Computer: YES, THREE OF THEM: A LARGE RED ONE, A LARGE GREEN CUBE, AND THE BLUE ONE.

Definite descriptions

Recognition of referential ambiguity

Confirmation

Pronouns

Negation

Can analyze state of the world to answer questions

Relative clauses

Indefinite description

Parsing

analyzing a sentence into its **constituents** (or parts) and describing their **syntactic roles** (parts of speech).

- grouping the words that 'go together'
- a parse **resolves** syntactic ambiguity
- one parse \Leftrightarrow one reading

e.g.

pancakes and bacon or eggs	\rightarrow	(pancakes and bacon) or eggs	PARSE 1
	\rightarrow	pancakes and (bacon or eggs)	PARSE 2

Parsing

What does this sentence mean?



[<http://www.clker.com/clipart-green-eyes-3.html>]



[<http://www.clker.com/clipart-3163.html>]



[<http://www.simonpalfrader.com/category/tournament-poker>]

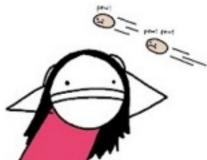
I **saw** her duck **with** a telescope.

(The word 'saw' is highlighted with a pink rectangle.)

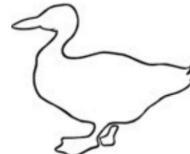
(The word 'saw' is highlighted with a pink rectangle.)

verb

noun



[<http://casablancape.blogspot.com/2010/05/fore.htm>]



[<http://www.supercoloring.com/pages/duck-outline/>]

How many **parses** are there for this sentence?

hint: how many **readings** are there for this sentence?

Parsing

S = Sentence

N = Noun

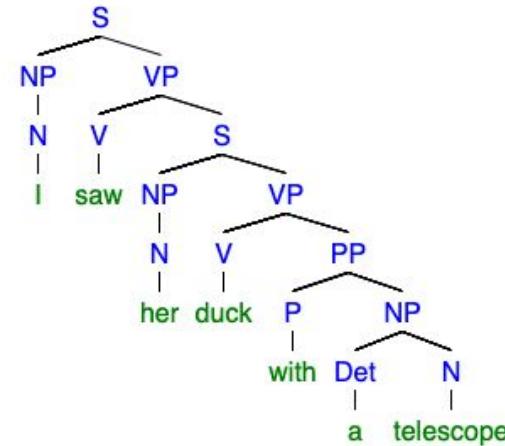
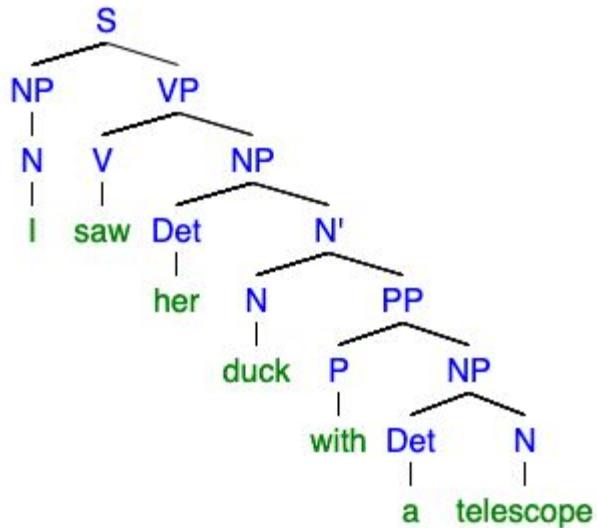
P = Preposition

V = Verb

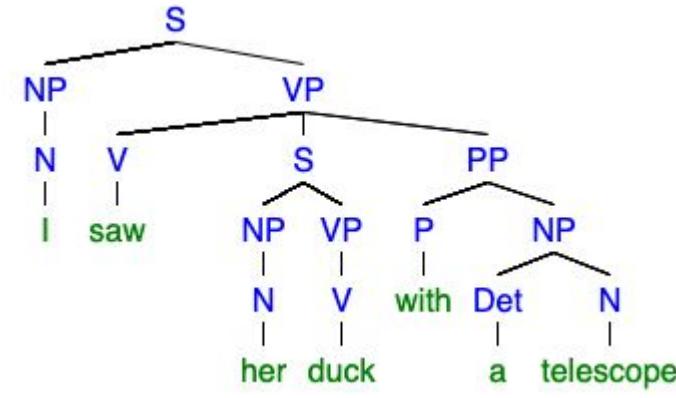
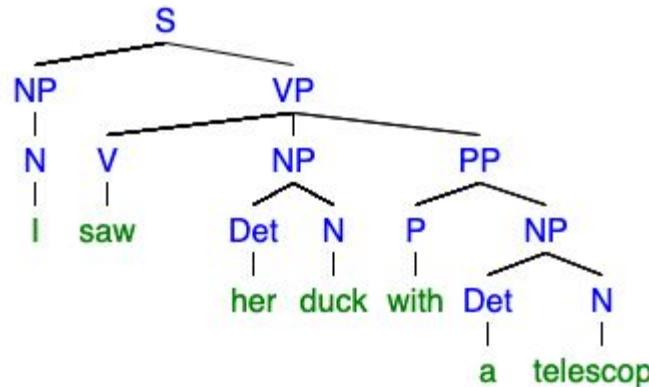
Det = Determiner
(aka article)

NP = Noun phrase

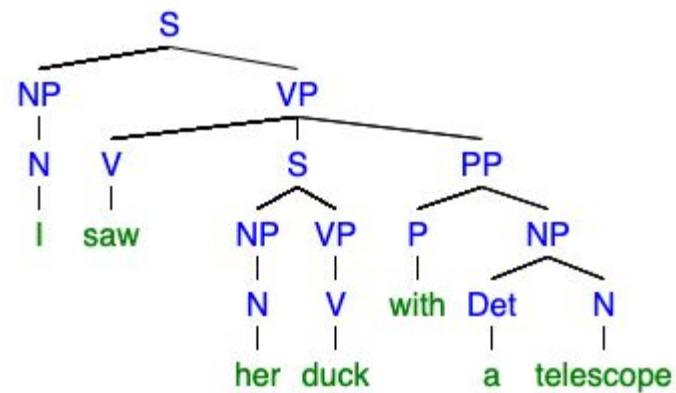
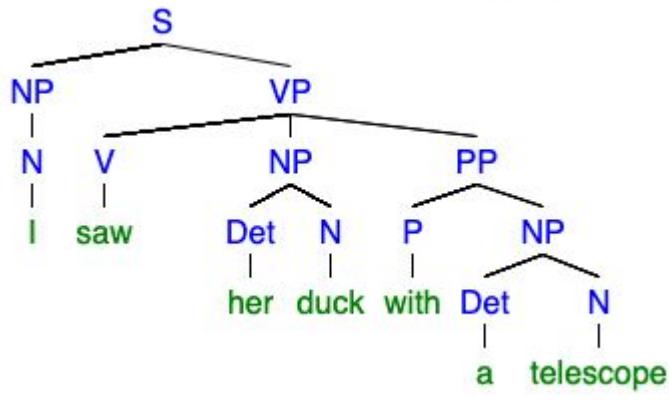
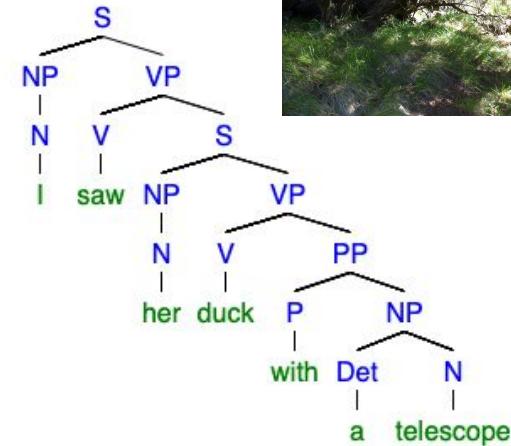
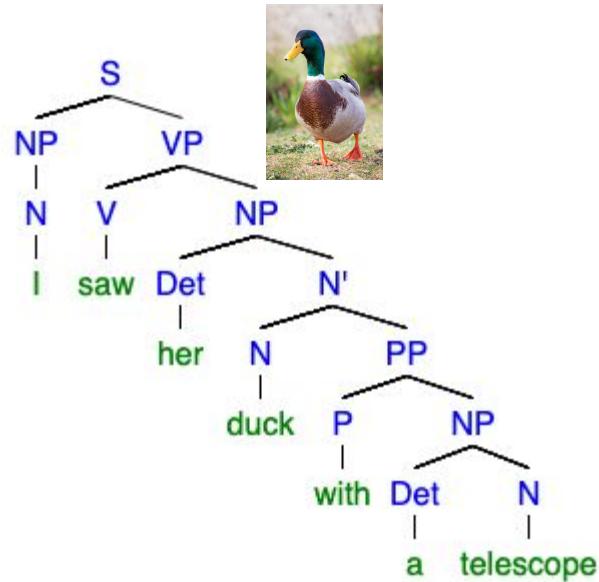
VP = Verb Phrase



Which trees correspond to the pet bird reading? Which trees are the ones where "I" have the telescope?



Parsing



- ambiguity can quickly get out of hand. how do we choose between all of the possible parses of a sentence?
- SHRDLU approach is inspired by observations about how humans process language. We seem to use meaning (semantics) to help us figure out how to parse a sentence, even before we have heard the whole thing
- That is, we don't wait until we've heard the whole sentence to *resolve* the ambiguity by selecting a meaning.
- Remember Chomsky's 'colorless green ideas' evidence that syntax and semantics are separate. There is also evidence that if they are separate, they are constantly interacting.

Understanding Natural Language (1972, Winograd's doctoral thesis): syntax and semantics work together in SHRDLU

The language-understanding problem is written in LISP to run under the PDP-10 ITS time-sharing system at the Artificial Intelligence Laboratory at MIT. The program is organized as indicated in Fig. 1. (Arrows indicate that one part of the program calls another directly.)

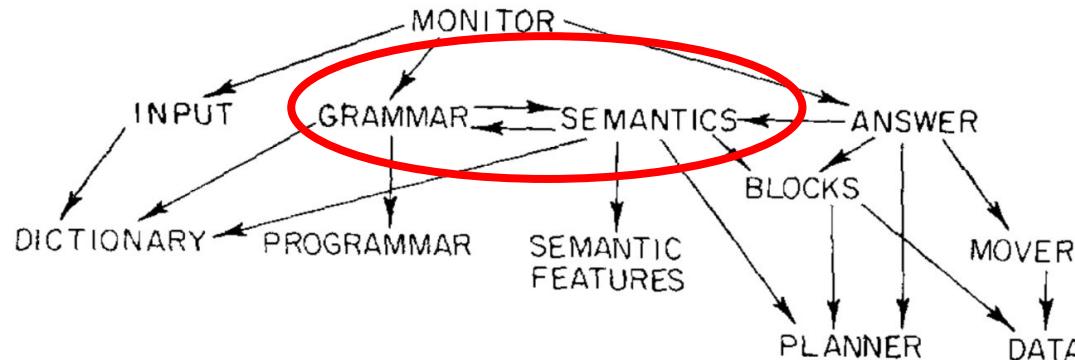


FIG. 1. Organization of the System.

[https://teaching.bb-ai.net/Student-Projects/Winograd-Challenge-Papers/
Winograd-1972-Understanding-Natural-Language.pdf](https://teaching.bb-ai.net/Student-Projects/Winograd-Challenge-Papers/Winograd-1972-Understanding-Natural-Language.pdf)

Understanding Natural Language (1972, Winograd's doctoral thesis): handling ambiguity

operation of a semantic interpreter. Rather, the process is unified, with the results of semantic interpretation being used to guide the parsing. The last statement in a program for parsing a noun group may be a call to a noun-group semantic interpreter. If it is impossible to interpret the phrase semantically at the time it is found, the parsing is immediately redirected.

The way of treating ambiguity is not through listing all possible interpretations of a sentence, but in being intelligent in looking for the first one, and being even more intelligent in looking for the next one if that fails. There is no automatic backup mechanism in PROGRAMMAR, because blind automatic backup is tremendously inefficient. A good PROGRAMMAR program will check itself when a failure occurs, and based on the structures it has seen and the reasons for the failure, it will decide specifically what should be tried next. PROGRAMMAR contains primitives for passing along reasons for failure and for performing the specific backup steps necessary.

Interlude: Garden Path Effects

The SHRDLU approach to handling ambiguity is

A **garden-path sentence** is a grammatically correct sentence that starts in such a way that a reader's most likely interpretation will be incorrect; the reader is lured into a parse that turns out to be a dead end or yields a clearly unintended meaning.

Garden Path Effects

A **garden-path sentence** is a grammatically correct sentence that starts in such a way that a reader's most likely interpretation will be incorrect; the reader is lured into a parse that turns out to be a dead end or yields a clearly unintended meaning.

e.g.,

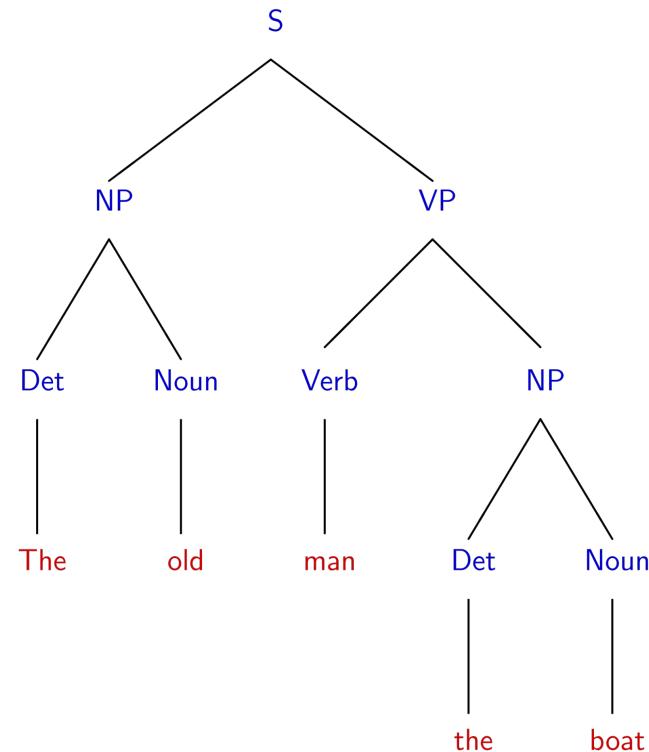
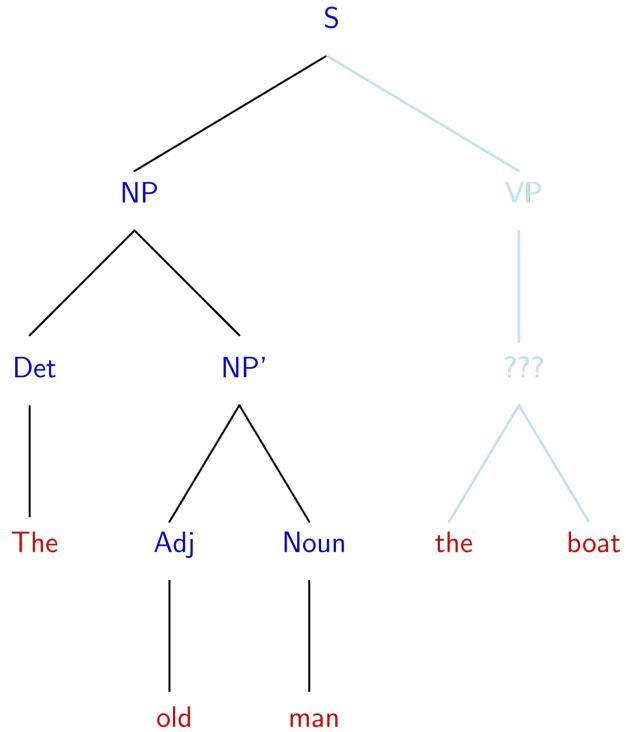
- (1) The old man the boat.
- (2) The horse raced past the barn fell.

Local vs Global parsing

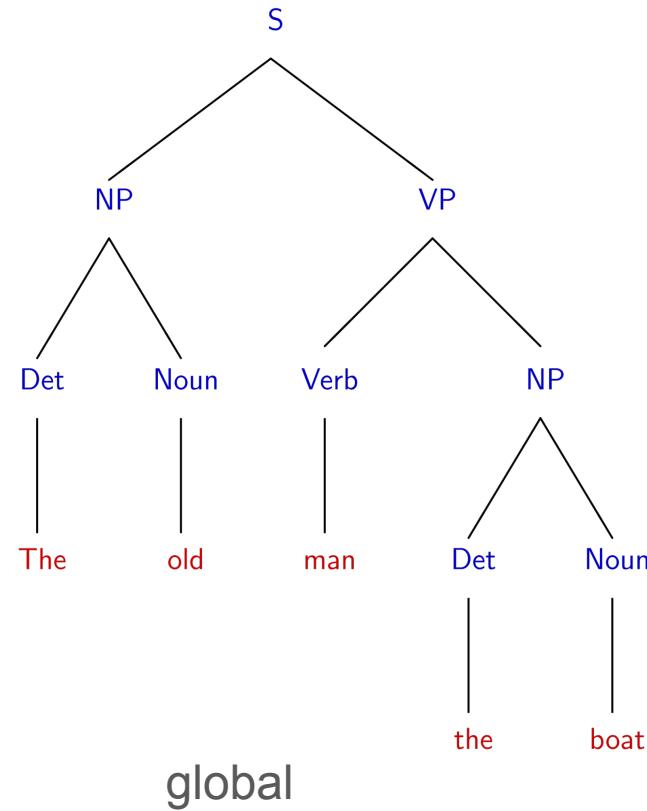
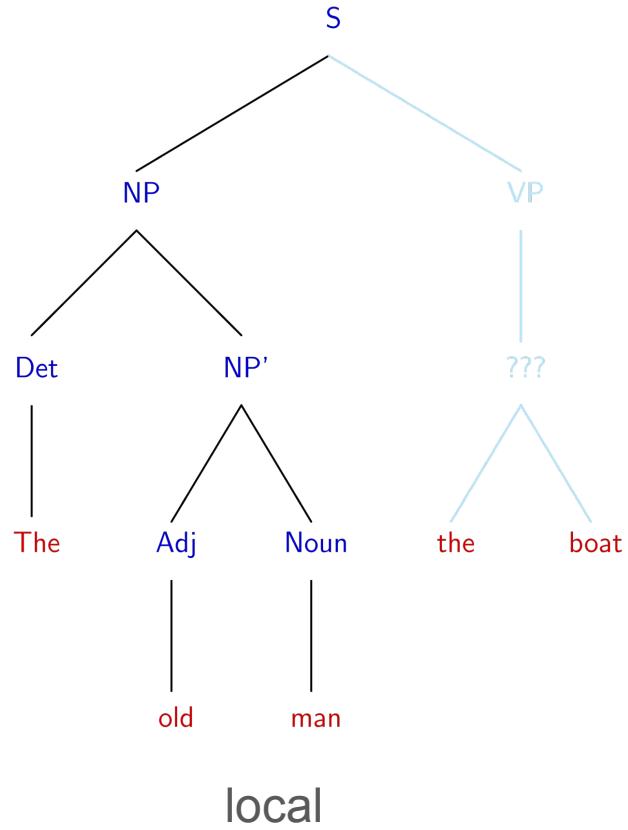
Garden path effects are evidence that we process sentences incrementally, looking at local information as we go along.

Even when a sentence is ambiguous, we don't wait until the end to resolve the ambiguity at a global level. The syntactic processor makes rapid decisions about which candidate is most likely.

Local vs Global parsing



Local vs Global parsing



Local vs Global parsing

Garden path effects are evidence that we process sentences incrementally, looking at local information as we go along.

Even when a sentence is ambiguous, we don't wait until the end to resolve the ambiguity at a global level. The syntactic processor makes rapid decisions about which candidate is most likely.

What information might the processor use to make such decisions?

A landscape painting featuring a winding path in the foreground leading towards a garden. The garden is filled with various flowers, including red poppies and yellow tulips. In the background, there are trees and a building with a tiled roof.

THE OLD MAN THE BOAT.

Local vs Global parsing

Garden path effects are evidence that we process sentences incrementally, looking at local information as we go along.

Even when a sentence is ambiguous, we don't wait until the end to resolve the ambiguity at a global level. The syntactic processor makes rapid decisions about which candidate is most likely.

What information might the processor use to make such decisions?

- SHRDLU: "more symbol knowledge"
- us: probabilities!

SHRDLU (*Understanding Natural Language*, Winograd 1972)

can be performed before going on. These may themselves do bits of parsing, or they may change the course the basic program will take after they are done. When we see the sentence “He gave the boy plants to water.” we don’t get tangled up in an interpretation which would be parallel to “He gave the house plants to charity.” The phrase “boy plants” doesn’t make sense like “house plants” or “boy scouts”, so we reject any parsing which would use it. This ability to integrate semantics with syntax is particularly important in handling discourse, where the interpretation of a sentence containing such things as pronouns may depend in complex ways on the preceding discourse and knowledge of the subject matter.

So...we need to know the possible relationships between everything and everything else?? Even the smallest problems become **intractable!** There are too many relationships to keep track of! And what about productivity? We can infer relationships even if we don't have prior knowledge or experience about something!

Referential Ambiguity: Winograd Schemas

Consider the following example from Terry Winograd (1972). Who does "they refer to"?

- (a) "The city councilmen refused the demonstrators a permit because **they** feared violence."

- (b) "The city councilmen refused the demonstrators a permit because **they** wanted violence."

Problems for Symbolic AI

combinatorial explosion

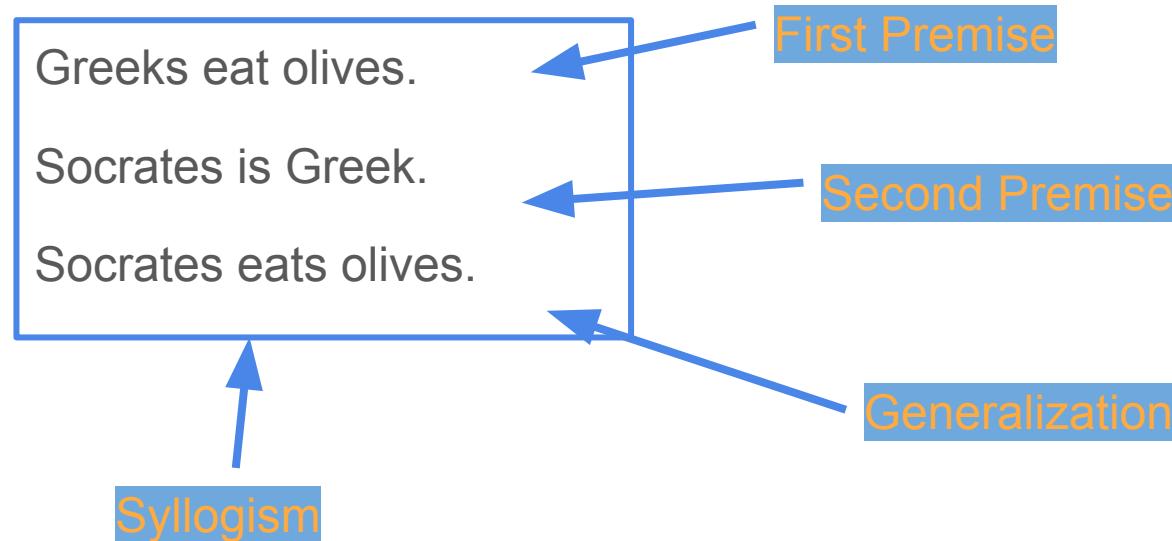
intractability

ambiguity (still)

Induction: Aristotle's **other** kind of reasoning

“reasoning from particulars to universals”

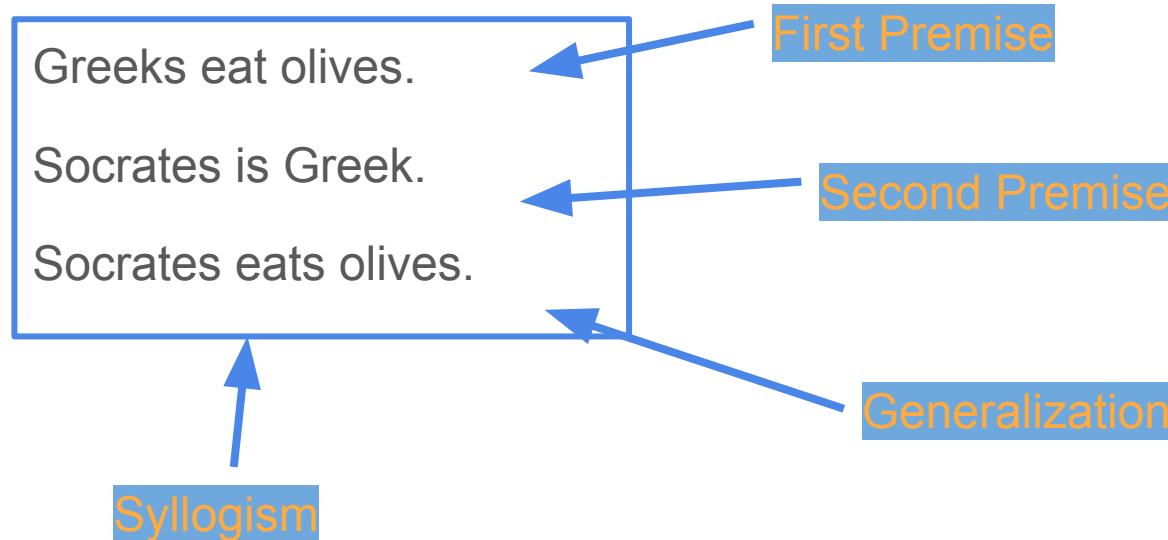
The conclusion is *likely* true based on the premises.



Induction: Aristotle's **other** kind of reasoning

“reasoning from particulars to universals”

The conclusion is *likely* true based on the premises.



The conclusion in inductive reasoning is **likely true** (versus definitely true for deductive syllogisms)

People can do both

We (people) use **deductive and inductive** reasoning:

Deductive reasoning: I put the forks in the drawer. Where are the forks?

Inductive reasoning: It rained every afternoon this week. Will it rain this afternoon?

Early AI researchers were well aware of this and had a **lot of feelings** about which one was more important for simulating intelligence

People can do both

We (people) use **deductive and inductive** reasoning:

Deductive reasoning: I put the forks in the drawer. Where are the forks?

Answer: yes

Inductive reasoning: It rained every afternoon this week. Will it rain this afternoon?

Answer: Probably

Early AI researchers were well aware of this and had a **lot of feelings** about which one was more important for simulating intelligence

There are two kinds of programs in this world...

- Option 1: Rule Based. Humans write instructions to make decisions
 - “If the word follows the definite article (“the”) and precedes a noun, label it as an adjective”
 - “If *apple* is capitalized outside of the first word in a sentence, label it as a company”
 - “If you see the word *excellent* in a review, label it as positive”
- Option 2: Machine Learning. Program figures out how to write instructions
 - Unsupervised Learning: figuring out patterns in data
 - Supervised Learning: learning by example

Subsymbolic AI

What if we don't need to *tell* the computer what the rules are ahead of time?

Introducing...the perceptron

FIG. 1 — Organization of a biological brain. (Red areas indicate active cells, responding to the letter X.)

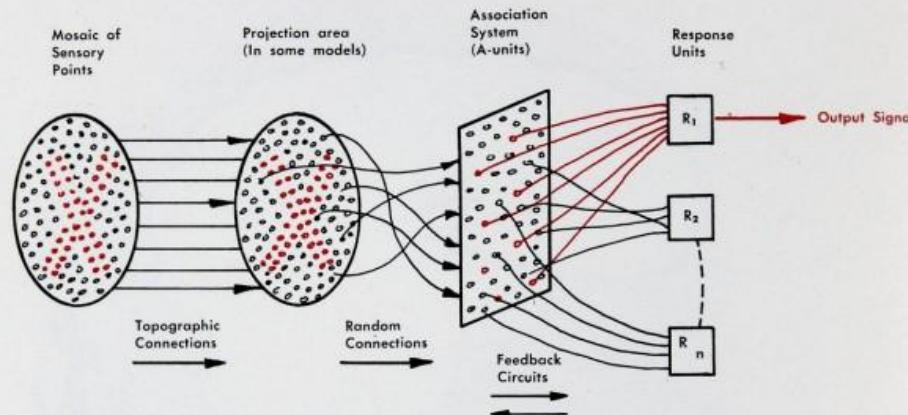
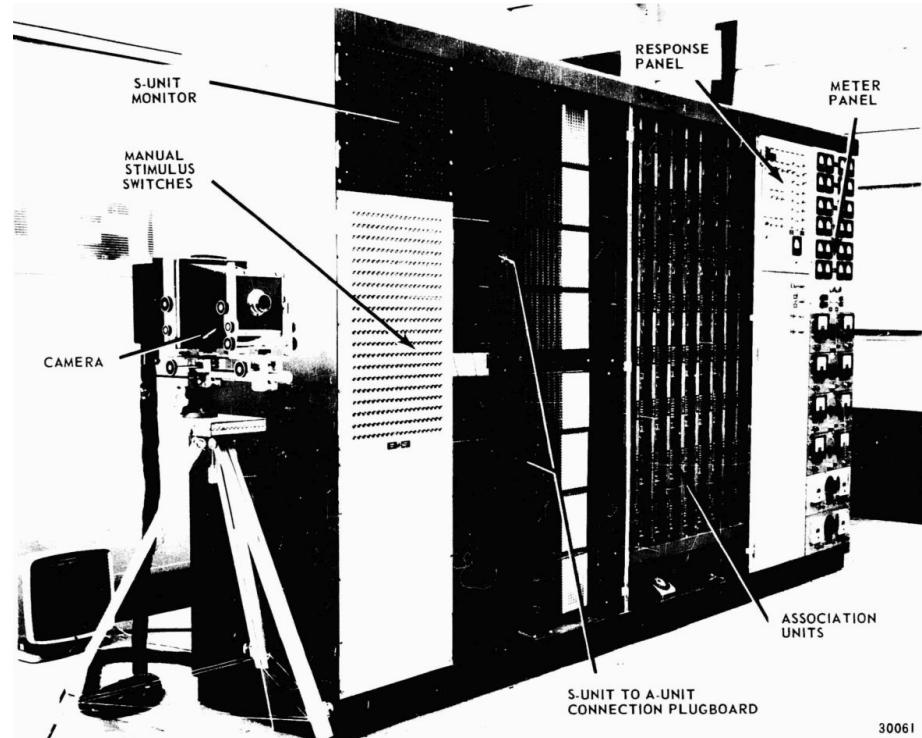


FIG. 2 — Organization of a perceptron.

Frank Rosenblatt (1958)

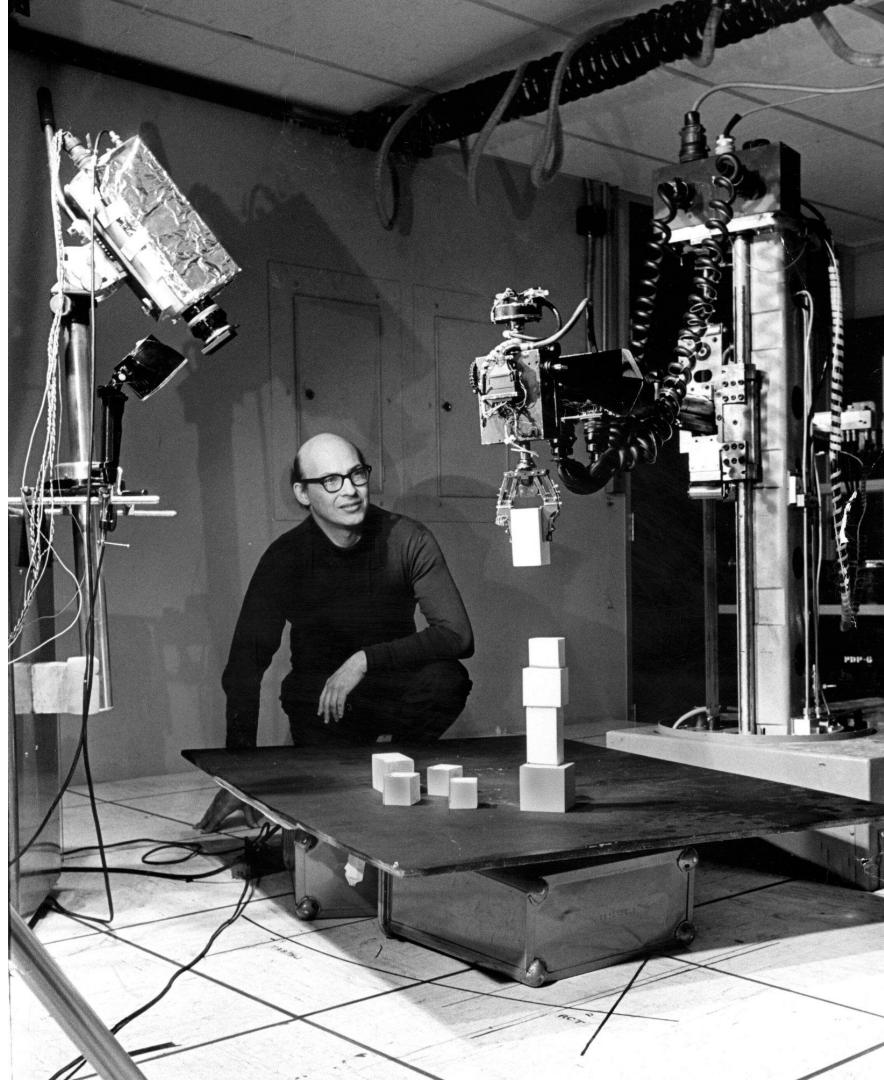
"Devices of this sort are expected ultimately to be capable of concept formation, language translation, collation of military intelligence, and the solution of problems through inductive logic."

- developed a *learning algorithm* for the perceptron
- builds an image classifier
- able to decide whether an X is on the left or right side of a page
- takes all day to train



And his rival, Marvin Minsky

- Big into symbolic AI
- original member of the Dartmouth summer research proposal
- Went to the Bronx High School of Science with Rosenblatt
- Pictured here stacking blocks



The inductive reasoning team is having a PR problem

Marvin Minsky and Seymour Papert wrote a book trashing perceptrons

- "[The perceptron] has many features to attract attention: its linearity; its intriguing learning theorem; its clear paradigmatic simplicity as a kind of parallel computation. There is no reason to suppose that any of these virtues carry over to the many-layered version. Nevertheless, **we consider it to be an important research problem to elucidate (or reject) our intuitive judgment that the extension is sterile**" *Perceptrons* (1969)

Argued for fundamental limitations to the Perceptron algorithm

First AI Winter (1974-1980)

Research funding dries up and public enthusiasm wanes

- They overpromised and underdelivered
- The Perceptron Controversy dampened enthusiasm for connectionist approaches
- Machine Translation Failures: The ALPAC report (1966) concluded that MT was more expensive and less accurate than human translation, leading to massive funding cuts
- Combinatorial Explosion: they work on toy problems but fail catastrophically on real-world complexity
 - the world is way bigger than blocks
 - language is bigger than sentences!
- Expert System Limitations: Rule-based systems are brittle and can't handle uncertainty or common-sense reasoning
 - even when we are talking about blocks

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale

Keisuke Sakaguchi*, Ronan Le Bras*, Chandra Bhagavatula*, Yejin Choi*†

*Allen Institute for Artificial Intelligence †University of Washington

{keisukes, ronanlb, chandrab, yejinc}@allenai.org

Abstract

The Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011), a benchmark for commonsense reasoning, is a set of 273 expert-crafted pronoun resolution problems originally designed to be unsolvable for statistical models that rely on selectional preferences or word associations. However, recent advances in neural language models have already reached around 90% accuracy on variants of WSC. This raises an important question whether these models have truly acquired robust commonsense capabilities or whether they rely on spurious biases in the datasets that lead to an overestimation of the true capabilities of machine commonsense.

To investigate this question, we introduce **WINOGRANDE**, a large-scale dataset of 4.4 million examples designed for the eval-

rely on statistical patterns without true capabilities of commonsense reasoning. However, recent advances in neural language models have already reported around 90% accuracy on a variant of WSC dataset.[†] This raises an important question:

Have neural language models successfully acquired commonsense or are we overestimating the true capabilities of machine commonsense?

This question about the potential overestimation leads to another crucial question regarding potential unwanted biases that the large-scale neural language models might be exploiting, essentially solving the problems *right*, but for *wrong* reasons. While WSC questions are expert-crafted, recent studies

Winogrande (2019)

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale

Keisuke Sakaguchi*, Ronan Le Bras*, Chandra Bhagavatula*, Yejin Choi*†

*Allen Institute for Artificial Intelligence †University of Washington

{keisukes, ronanlb, chandrab, yejinc}@allenai.org

Abstract

The Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011), a benchmark for commonsense reasoning, is a set of 273 expert-crafted pronoun resolution problems originally designed to be unsolvable for statistical models that rely on selectional preferences or word associations. However, recent advances in neural language models have already reached around 90% accuracy on variants of WSC. This raises an important question whether these models have truly acquired robust commonsense capabilities or whether they rely on spurious biases in the datasets that lead to an overestimation of the true capabilities of machine commonsense.

To investigate this question, we introduce WINOGRANDE, a large-scale dataset of 6411 problems. Interestingly, the main

rely on statistical patterns without true capabilities of commonsense reasoning. However, recent advances in neural language models have already reported around 90% accuracy on a variant of WSC dataset.[†] This raises an important question:

Have neural language models successfully acquired commonsense or are we overestimating the true capabilities of machine commonsense?

This question about the potential overestimation leads to another crucial question regarding potential unwanted biases that the large-scale neural language models might be exploiting, essentially solving the problems *right*, but for *wrong* reasons. While WSC questions are expert-crafted, recent studies

Winogrande (2019)

Twin sentences

✗ The monkey **loved** to play with the balls but **ignored** the blocks because he found **them** **exciting**.

The monkey **loved** to play with the balls but **ignored** the blocks because he found **them** **dull**.

✗ William could only climb **begginer** walls while Jason climbed advanced ones because **he** was very **weak**.

William could only climb **begginer** walls while Jason climbed advanced ones because **he** was very **strong**.

✓ Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had **less** time to get ready for school.

Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had **more** time to get ready for school.

✓ The child was screaming after the baby bottle and toy fell. Since the child was **hungry**, **it** stopped his crying.

The child was screaming after the baby bottle and toy fell. Since the child was **full**, **it** stopped his crying.

Winogrande

Twin sentences		Options (answer)
<input checked="" type="checkbox"/> The monkey loved to play with the balls but ignored the blocks because he found them exciting.		balls / blocks
The monkey loved to play with the balls but ignored the blocks because he found them dull.		balls / blocks
<input checked="" type="checkbox"/> William could only climb beginner walls while Jason climbed advanced ones because he was very weak.		William / Jason
William could only climb beginner walls while Jason climbed advanced ones because he was very strong.		William / Jason
<input checked="" type="checkbox"/> Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had less time to get ready for school.		Robert / Samuel
Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had more time to get ready for school.		Robert / Samuel
<input checked="" type="checkbox"/> The child was screaming after the baby bottle and toy fell. Since the child was hungry, it stopped his crying.		baby bottle / toy
The child was screaming after the baby bottle and toy fell. Since the child was full, it stopped his crying.		baby bottle / toy

Winogrande

Twin sentences

Options (answer)

The monkey loved to play with the balls but ignored the blocks because he found them exciting.

balls / blocks

The monkey loved to play with the balls but ignored the blocks because he found them dull.

balls / blocks

William could only climb beginner walls while Jason climbed advanced ones because he was very weak.

William / Jason

William could only climb beginner walls while Jason climbed advanced ones because he was very strong.

William / Jason

Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had less time to get ready for school.

Robert / Samuel

Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had more time to get ready for school.

Robert / Samuel

The child was screaming after the baby bottle and toy fell. Since the child was hungry, it stopped his crying.

baby bottle / toy

The child was screaming after the baby bottle and toy fell. Since the child was full, it stopped his crying.

baby bottle / toy

SOLVED?

What makes language hard (for language models) now?

What makes language hard (for language models) now?

- social cues
- common sense
- truth / world knowledge
- *action*
- ...?