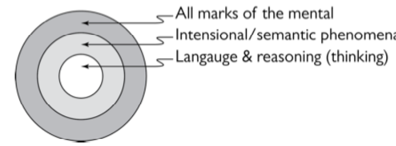
Part II — The Classical Model

Formal Representational and Logic

The Plot So Far



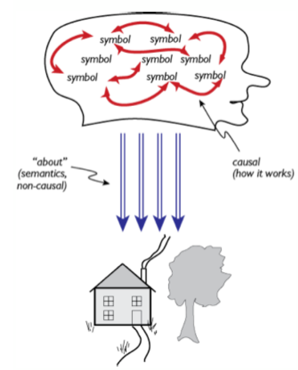
1. **Subject matter**
   a) *Narrowing #1:* Of many potentially distinguishing "marks of the mental," the ones that have been primarily focused on, in intellectual history—especially in AI, cognitive science, and philosophy of science—are **intentional/semantic** ones.
   b) *Narrowing #2:* Of all intentional/semantic phenomena, the ones that philosophy, AI and cognitive science have primarily studied, over the last 50 years, have been **language** and **thinking**.

The Plot So Far (cont'd)

2. **Method**
   a) In terms of *how to study* language and reasoning (thinking), 3 options initially presented themselves:
      i) Consider the mind **without internal states**—just look at "externally observable behaviour" (called behaviourism);
      ii) Study the **neural inner workings** of the brain, on the grounds that the brain is the seat of intelligence and mind (i.e., do neuroscience); or
      iii) Model the mind on *how we understand minds in "***folk psychology***"*—i.e., how we describe mental states when talking about other people.
   b) The first (behaviourism) didn't work.
   c) When AI/CogSci got started, the second was technically impossible. In addition—and these reasons remain compelling, even if it is becoming more and more technically possible—it seemed both too *low level* and too *"human chauvinist."*
   d) So people chose the 3rd: *folk psychology*, which led to the postulation of an an **internal representational language**—something often called "mentalese"—that is structurally similar to natural languages, with properties of **productivity**, **systematicity**, and **compositionality**.

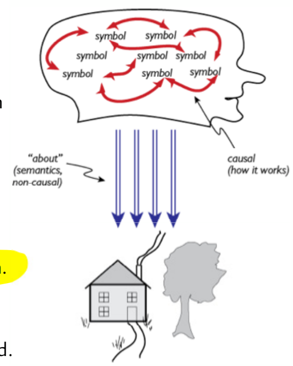
The Plot So Far (cont'd)

1. Mentalese expressions are taken to be **symbols**, which **represent** the world that the agent inhabits (that is: represent states of affairs that they think about).
2. Mentalese symbols participate in **two critical relations**, which must be coordinated
   a) **Causal** relationships, which transforms the symbols into other ones—the *how thinking works* part.
      — Indicated (in this course!) with single, red arrows
   b) A **semantic** relation of *aboutness*, which relates the symbols to what it is that they *represent* or *refer to*
      — Indicated (in this course!) with double, blue arrows



The Plot So Far (cont'd)

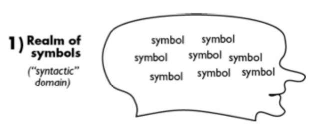
3. This fundamental structure applies to representational theories of both minds and machines:
   a) The system (mind or machine) must *work, causally*, in virtue of the causal relations (red arrows).
   b) To work **properly**, however, requires that the symbols in the system (mind or machine) **remain appropriately coordinated** with the world they refer to or are about (via the blue arrows).
4. This implies that all such systems are *governed by a* **norm**.
   a) That a representation be *true* is perhaps the most obvious instance of this norm; it is one example of how thoughts have to be coordinated with the world.
5. Without this coordinating norm, a representational system is nothing!
6. And—a fact that will be extremely important fact for us throughout the rest of the course—the *norm cannot be expressed in purely causal terms.*



Formal Representational and Logic

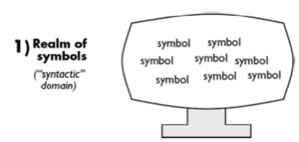
1. The classical model of AI—pretty much *all* of AI, in its original decades, now called (by John Haugeland) "**GOFAI**", for "Good Old-Fashioned Artificial Intelligence"—is ultimately based on a **formal symbol manipulation (FSM)** model of *computation*
2. Formal symbol manipulation, in turn, is classically exemplified in **formal logic**
3. To understand GOFAI, therefore, we need to understand formal logic
4. Not (so much!) the technical details ...
   a) $[\forall x\,[\text{man}(x) \supset \text{mortal}(x)]] \wedge [\text{man}(\text{Socrates})] \supset \text{mortal}(\text{Socrates})]$
   b) $[P(x) \supset Q(x)] \Leftrightarrow [\neg Q(x) \vee P(x)]$
   c) $\{A, A \supset C, \neg C\} \vdash \neg A$
   d) $w_0 \vDash \forall s \mid [\text{set}(s) \wedge \forall p \mid [\, p \in s \supset [\text{person}(p)]]]\ [\exists p' \in s \mid \text{sad}\,(p')]$
5. But the **fundamental idea** behind the formal logic tradition
6. What *is* the fundamental idea behind formal logic?
7. It is *nothing other than a formalization of the representational theory we have just reviewed!*

The structure of logic and logic-like systems



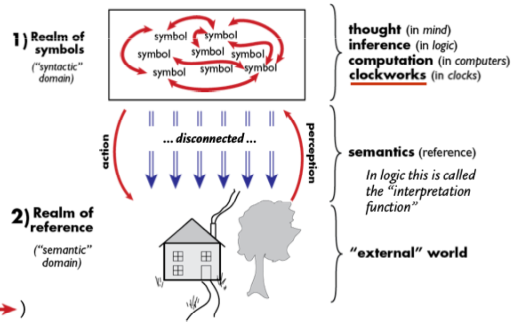
1. Two domains
   ✔ a) A realm of **symbols** ("syntactic domain")
   b) A realm of **reference** ("semantic domain")

The structure of logic and logic-like systems



1. Two domains
   ✔ a) A realm of **symbols** ("syntactic domain")
   b) A realm of **reference** ("semantic domain")

The structure of logic and logic-like systems



1. Two **domains**
   - ✔ a) A realm of **symbols** ("syntactic domain")
   - ✔ b) A realm of **reference** ("semantic domain")
2. Two kinds of **relationship** involving the two realms
   - ✔ a) A causal/mechanical relationship of **inference** (→)
   - ✔ b) An intentional relationship of **semantics** (⇒)

*In cognitive science & computing (though not logic per se) there are more relationships, which are both causal/mechanical and also intentional/semantic, connecting the system to the (external) world—***action** *&* **perception**.

3. A **normative constraint**, that
   - a) Holds among the domains and relationships
   - b) Applies to one of the relationships

In general, the norms mandate **coordination**: the transitions among the symbols must *make sense*, in terms of what they represent or refer to. Overall, the symbols must **defer** to the worlds they represent.

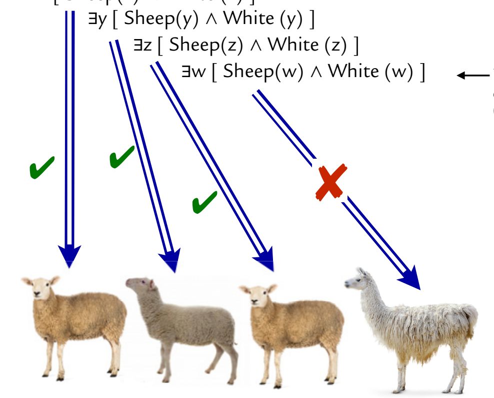
Realm of Symbols

∃x [Sheep(x) ∧ White (x)]
∃y [Sheep(y) ∧ White (y)]
∃z [Sheep(z) ∧ White (z)]
∃w [Sheep(w) ∧ White (w)]

← *This* is failure of statement (it is *false*)

Realm of Semantics

Realm of Reference



Realm of Symbols

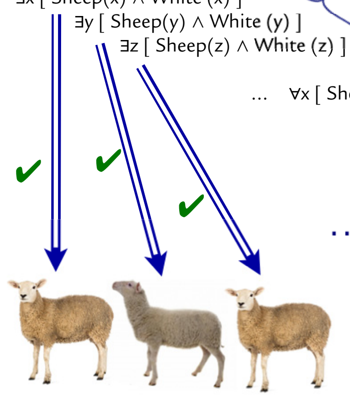
∃x [Sheep(x) ∧ White (x)]
∃y [Sheep(y) ∧ White (y)]
∃z [Sheep(z) ∧ White (z)]

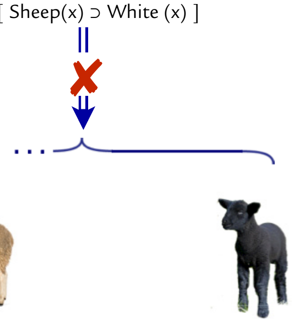
This is not a legitimate conclusion of *these* (so this is a failure of *reasoning*)

... ∀x [Sheep(x) ⊃ White (x)]

Realm of Semantics

Realm of Reference





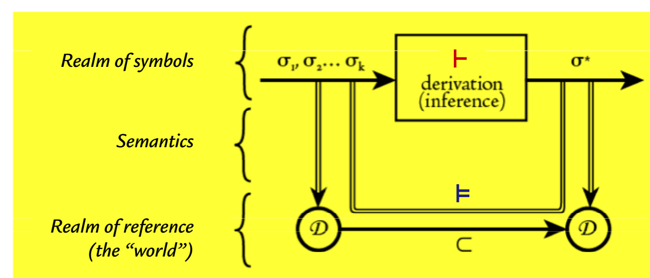
The nature & role of these **norms** can be understood more by looking at additional cases:

1. There are norms for action: that what is *achieved* is what was *intended;*
2. For perception: that the symbols [or thoughts] caused by perception *represent their cause*

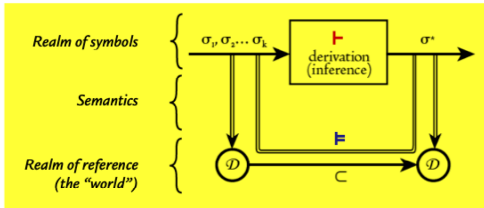
... etc.

But let's look at the case that has been worked out in most detail: that of *formal logic*

Logic, closer to way in which it is normally presented



Logical inference (and proof theory)



Proof (inference):

causal (*formal*, syntactic, effective)

$\{\{P \supset Q\}, \neg Q\} \vdash \neg P$

$\{[\forall x\,[\mathrm{man}(x) \supset \mathrm{mortal}(x)]], [\mathrm{man}(\mathrm{socrates})]\} \vdash \mathrm{mortal}(\mathrm{socrates})$

(All men are mortal; Socrates is a man. Therefore: Socrates is mortal)

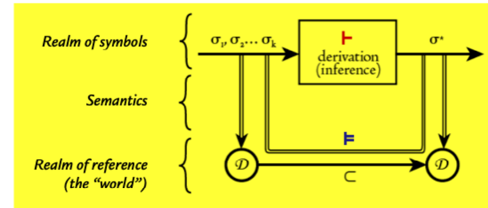
Entailment:

$\{Si\} \vDash \forall s \mid [\mathrm{set}(s) \wedge \forall p \mid [\, p \in s \supset [\mathrm{person}(p)]]]\ [\exists p' \in s \mid \mathrm{sad}\,(p')]$

(In any set of people, there will always be someone who is sad)

semantic (non-causal, non-effective)

The 5th ingredient: the fundamental normative constraint



1. Logic (and symbol manipulation) should be **truth-preserving**
   — *If the premises are true, the conclusion should be true, where whether a sentence is true is a semantic property of the sentence, not a syntactic or formal property.*
2. **Syntax** should honour **semantics**
3. $\rightarrow$ should honour $\Rightarrow$
4. $\vdash$ should honour $\vDash$

Criteria on logical systems, based on the constraint

1. Soundness: **wanting what you get**
   — i.e., $\vdash$ *only when* $\vDash$
2. Completeness: **getting what you want**
   — i.e., $\vdash$ *whenever* $\vDash$

Another case, maybe easier to understand: *Clocks*

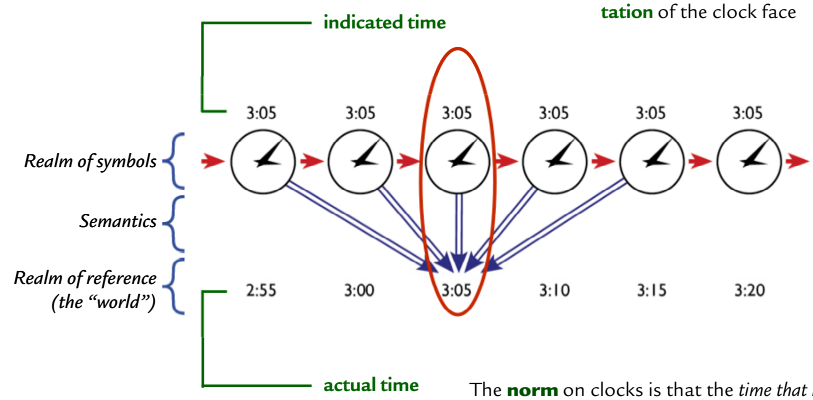
The world's only perfectly accurate clock:

The **norm** on clocks is that the *time that is indicated* should be the *time that it is*.

A clock that is stopped

$\rightarrow$ is the **clockworks**

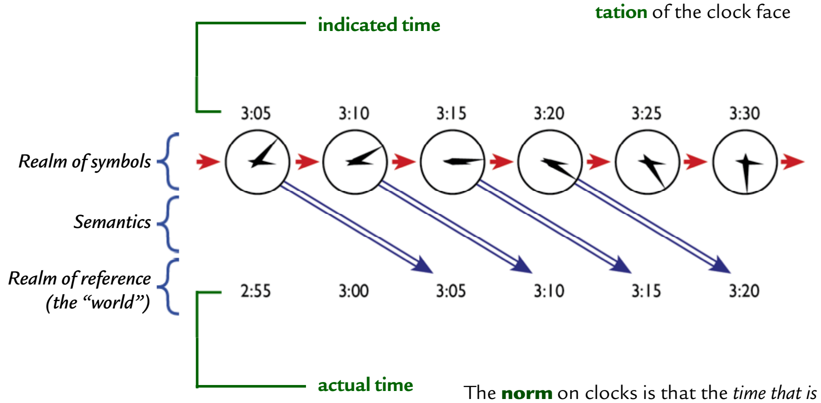
$\Rightarrow$ is your **semantic interpretation** of the clock face



The **norm** on clocks is that the *time that is indicated* should be the *time that it is* (i.e., the semantic arrow should be *vertical*).

A clock that is running 10 minutes late

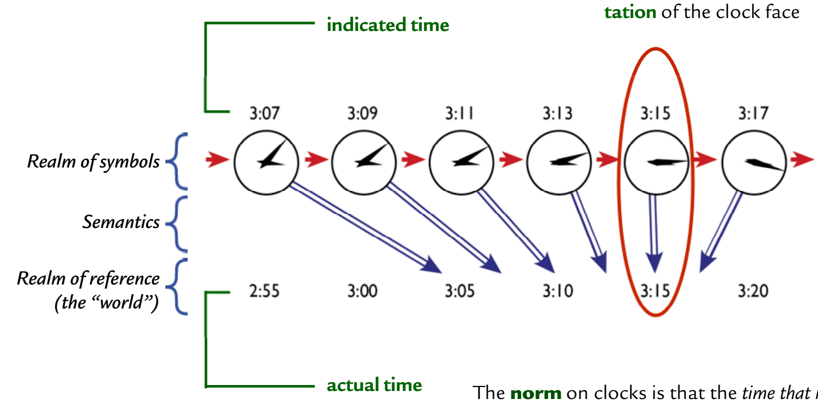
is the **clockworks**
is your **semantic interpretation** of the clock face



The **norm** on clocks is that the *time that is indicated* should be the *time that it is* (i.e., the semantic arrow should be *vertical*).

A clock that is running slowly

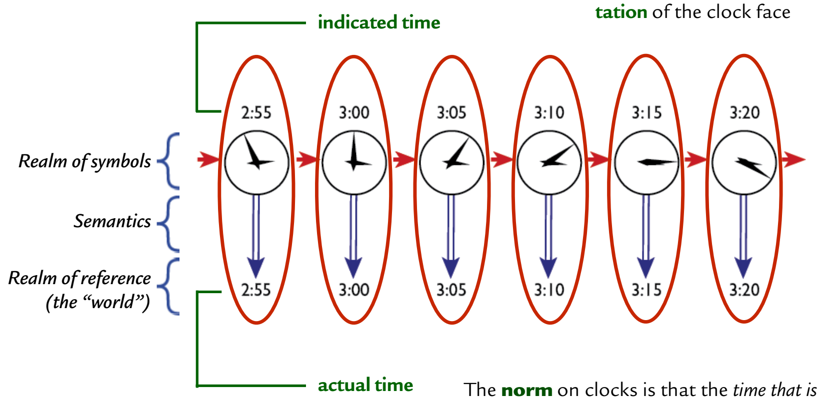
is the **clockworks**
is your **semantic interpretation** of the clock face



The **norm** on clocks is that the *time that is indicated* should be the *time that it is* (i.e., the semantic arrow should be *vertical*).

A clock that is correct (and running correctly)

is the **clockworks**
is your **semantic interpretation** of the clock face



The **norm** on clocks is that the *time that is indicated* should be the *time that it is* (i.e., the semantic arrow should be *vertical*).

Formally (if you are into such things):

$\frac{\partial}{\partial t}$ $[\![$  $]\!]$ $= 1$

The Clock Theorem

The Plot for Today

1. At a fundamental level, all representational systems exhibit the same structure:
   a) A set of causally efficacious ingredients or states (symbols, expressions, mental states, computational states, states of mechanism [e.g., states of clockworks], etc.);
   b) An external world or task domain;
   c) A semantic interpretation ($\Longrightarrow$) of those ingredients or states, which says what the ingredients or states refer to or are "about";
   d) A causal mechanism ($\longrightarrow$) defined in terms of the causally efficacious ingredients (inference in the case of logic, thinking in the case of the mind, "computing" in the case of computers, "running" in the case of the clock, etc.); and
   e) A norm that places conditions on the ingredients or states, and on the causal mechanism, in order to ensure appropriate coordination with the world to which the ingredients or states are semantically related (towards which they "semantically point")
2. Though the details are different in the various different cases (mind, computers, logic, clocks, etc.), the fundamental architecture is the same.
3. This is what we will be calling the "**classical model**."

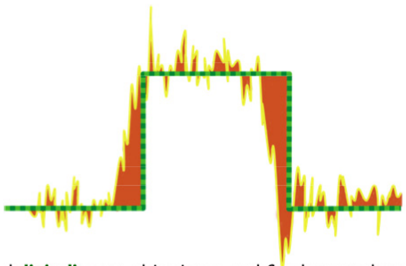
One more thing ...

The "standards" on mind

1. Keep in mind how high a standard we are setting, in terms of what it is to be a *mind*, or to be *mental*.
2. An impressive non-example (i.e., something that Descartes would *not* consider as requiring a mind, and therefore a behaviour of a "mere beast"): David Gallo's **Octopus Intelligence***

www.youtube.com/watch?v=PmDTtkZlMwM

Formality and Digitality



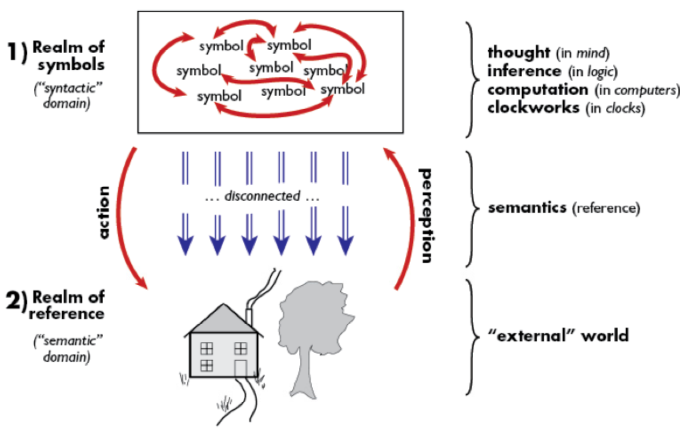
Formality and **digitality** are ubiquitous and fundamental notions that:

1. Underlie (virtually all of) computation,
2. Have played a huge role in revolutionizing contemporary life, and
3. Are basic assumptions of the classical "GOFAI" model of AI and cognitive science,
4. … but are *rarely analyzed in and of themselves!*

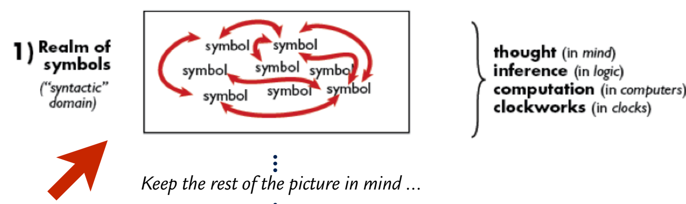
GOFAI and the *Mechanism* of Mind

1. As we saw last Tuesday, according to the classical "GOFAI" model, the **normative** conditions on mind arise from the **semantical** (⟹) side of things—from the fact that mental states are *about* states of affairs in the world of the mind or machine.
2. Today, however, and for the next while, we will focus on the *causal/effective/mechanical* side (⟶): on the **nature of the mechanism** that "implements" the mind.
3. Because we are talking about GOFAI, we will start with Haugeland's characterization of the mechanical conditions on **formal symbol manipulation**.
   a) Cf. Haugeland's "Semantic Engines" and "Automatic Formal Systems" (on Blackboard)
4. Other sorts of mechanism will be examined in Part III of the course.

Haugeland's characterization can be understood in terms of the classical model:



Haugeland's characterization can be understood in terms of the classical model:



⋮
Keep the rest of the picture in mind …
⋮

1. Issues about the "**mind as mechanism**" (the realm of symbols)
2. What Haugeland calls an "**Automatic Formal System**"
3. A computer, he says—and hence the mind, according to GOFAI—is an:
   ✔ a) **interpreted** ⟸ i.e., has **semantics** (⟹)
   ✔ b) **automatic** ⟸ **runs on its own**
   c) **formal system** ⟸ **???**
4. We have talked about the first (semantic interpretation)
5. The second (automatic) is fairly clear
6. Today: look at the third: *formality*

what does this mean ?

Formality #1 — Negative (the usual definition)

1. The notion of *formality* applies to systems (like the ones we are studying) with both
   a) Causal or syntactic properties ($\longrightarrow$, $\vdash$) and
   b) Semantic properties ($\Longrightarrow$, $\models$)
2. A system is *formal*, according to the most standard definition, just in case *the causal (syntactic, mechanical) parts work independently of the semantic parts*
3. "**Syntax independent of semantics**," this is often put—especially in the context of formal logic.
4. The problem with this definition of formality (and thus of the category of formal systems) is that it is *negatively defined*. To know what it comes do, one would need to know what semantic properties are like—and that is an issue on which there is not much agreement (and perhaps not yet much illumination!)
5. Intuitively, on the other hand, there seems to be great deal of agreement on what formal systems are, which systems are formal, etc.
6. That raises the question of whether we can't formulate a *positive* definition of formality that captures people's intuitive sense.
7. That is what Haugeland has tried to do—to come up with a *positive* characterization of formality

Formality #2 — Positive (Haugeland)

computer — interpreted — with semantics ($\Longrightarrow$)
automatic — runs on its own
formal — token-manipulation
digital
finitely-playable

1. Token manipulation
2. Digital
3. Finitely-playable

Formality #2 — Positive (Haugeland)

computer — interpreted — with semantics ($\Longrightarrow$)
automatic — runs on its own
formal — token-manipulation
digital
finitely-playable

1. Token manipulation
   a. Cf. checkers, chess, tic-tac-toe
   b. Cf. 0's and 1's in a "computer"
   c. Self-contained (no "outside world")
   d. Must be physical (0s & 1s?)

   What's with that?
   Are 0s and 1s physical?

   i. Tokens of a **finite number of disjoint types** (that is: an *unbounded* number of potential tokens, all of which are instances of a *finite* number of types)
   ii. **Finite** rules (formulated in terms of the finite types)
   iii. **Unbounded** input (via recursion)
   iv. Gives the system a degree of **abstractness**
   v. Necessary for **medium independence**
2. Digital
3. Finitely-playable

Formality #2 — Positive (Haugeland)

computer — interpreted — with semantics ($\Longrightarrow$)
automatic — runs on its own
formal — token-manipulation
digital
finitely-playable

1. Token manipulation
2. Digital
3. Finitely-playable

Formality #2 — Positive (Haugeland)

- computer
  - interpreted — with semantics (⟶)
  - automatic — runs on its own
  - formal
    - token-manipulation
    - digital
    - finitely-playable

▶ 1. Token manipulation

▶ 2. Digital

▼ 3. Finitely-playable

a) Must be able to operate ("work") via a **finite mechanism**

b) No infinities (and no impractically large numbers, either, like examining *all possible sequences of moves in a chess game*, of which there are something like 10^{120})

c) No access to *non-effective* properties (such as what time it is!)

d) No answers to questions that only oracles can answer (such as whether intelligent aliens will visit us within the next 100 years, or whether a woman will be elected President of the U.S.A. before 2032).

e) In other words, something that can be done by a **practical machine!**

Formality #2 — Positive (Haugeland)

- computer
  - interpreted — with semantics (⟶)
  - automatic — runs on its own
  - formal
    - token-manipulation
    - digital
    - finitely-playable

▶ 1. Token manipulation

▶ 2. Digital

▶ 3. Finitely-playable

Formality #2 — Positive (Haugeland)

- computer
  - interpreted — with semantics (⟶)
  - automatic — runs on its own
  - formal
    - token-manipulation
    - digital
    - finitely-playable

▶ 1. Token manipulation

▼ 2. Digital

a) One of the most **consequential**, **profound**—and **poorly understood**—phenomena underlying the computer revolution!

b) In spite of its enormous popularity!

c) On Oct 3, 2016, a search for 'digital' on Amazon* produced:

i. 62,760,799 items

ii. 4,607,821 books

Yikes!

d) We need to look seriously at what this ubiquitous word 'digital' actually *means*

▶ 3. Finitely-playable

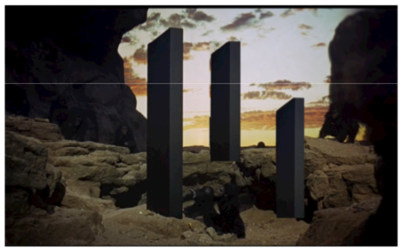
*On www.amazon.com. On the Canadian site (www.amazon.ca) the same search produced only 10,870,950 items and 4,325,072 books (presumably the best ones ;-)). On Sept 24, 2017, the situation seems different. The American site (www.amazon.com) no longer seems to say how many results a search yields. On the Canadian site, 'digital' produces 2,477,092 results if one searches all categories, but 5,086,011 if one restricts it to books. Go figure. But it is all meaningless anyway; the point is only that the term has permeated contemporary discourse and materiality.

Digitality

1. Everyone knows:
   a) That computers are digital (or most of them, anyway)
   b) That the invention of the digital computer was one of "the" inventions of all time
   c) That, somehow or other, the emergence of the digital computer gave us *abstract symbols*, *universal machines*, *programming languages*, *data bases*, *digital controllers*—and the *internet* ... to say nothing of *CDs* and *DVDs*, *personal computers*, *e-mail*, *mobile smartphones*, *electronic gaming*, *virtual reality* ... and so on.
2. What far fewer people know:
   a) What '**digital**' actually means
   b) What it is to be digital—what properties "**being digital**" conveys

Digitality (cont'd)

3. Tellingly, if you ask people what 'digital' means, they often answer *ostensively*—saying that a digital system is *like:*



Or, as *like the integers:*
1, 2, 3, 4, 5, 6, 7, …

Digitality (cont'd)

4. What would it be to *define* digitality—to say **what properties a system has to have**, in order to be a **digital** system?
5. Again, that is what Haugeland tries to do. I.e.,
   a) Just as he offers a **positive characterization of formality**
   b) So, too, he offers a **positive** (non-ostensive) **definition of digitality**

Digitality (cont'd)

So what is digitality? According to Haugeland, a digital system requires:

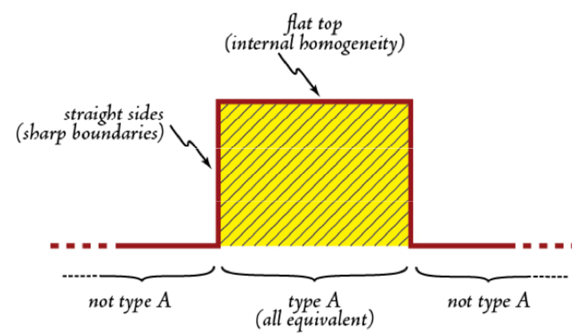
1. A set of distinct **types**
2. Each type must have a set of *absolutely identical, indistinguishable* (for purposes of the system) **tokens**
   a) Cf. checkers, chess, tic-tac-toe
   b) Cf. 0's and 1's in a "computer"

Haugeland's "token manipulation," which we saw above

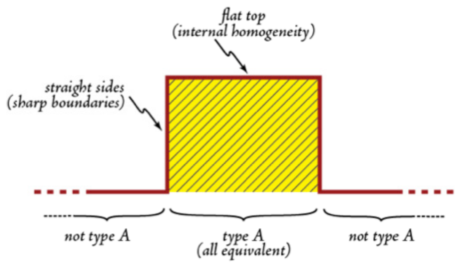
3. Questions must have **absolute, definite, yes/no answers**:
   a) "Is α a token of type β?"
   b) "What type is α a token of?"
4. No *ambiguity*, no *vagueness*, no *matters of degree*
5. I.e.: *perfect copyability*, *perfect reproducibility*, *absolute determination of types*, etc.
6. In other words: a **perfect** system of **utterly reliable interchangeable parts**

Eli Whitney to the max!

A (BCS) symbolic representation of this characterization of digitality



Problem

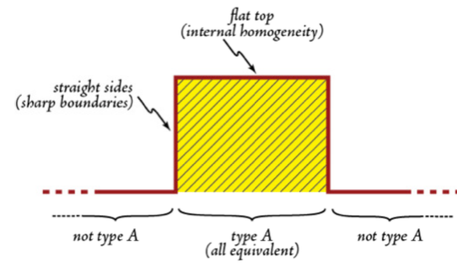


1. So far, this sounds like abstract mathematics
2. Discrete, perfect, types & tokens
3. What does this have to do with computing, and with the digital revolution?
4. In particular, how is anything like this—anything with this sort of "perfect or perfected clarity"—**possible** in the messy, disheveled world we live in—a world of friction, decay, sloppiness, etc.?

"Where moth and rust do corrupt"

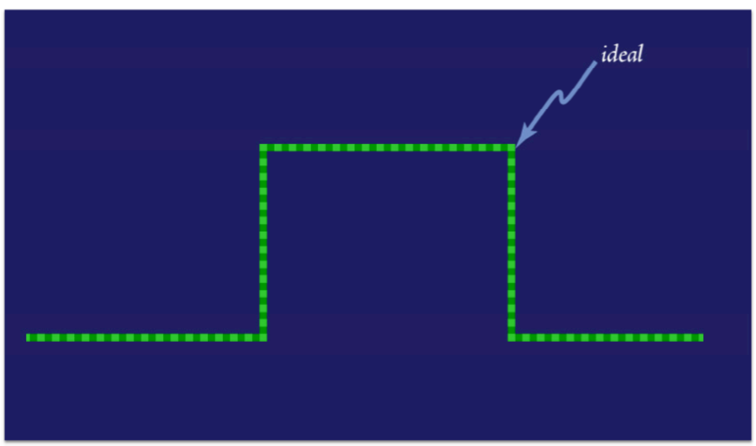
The $1,000,000,000,000 question!

Problem (cont'd)

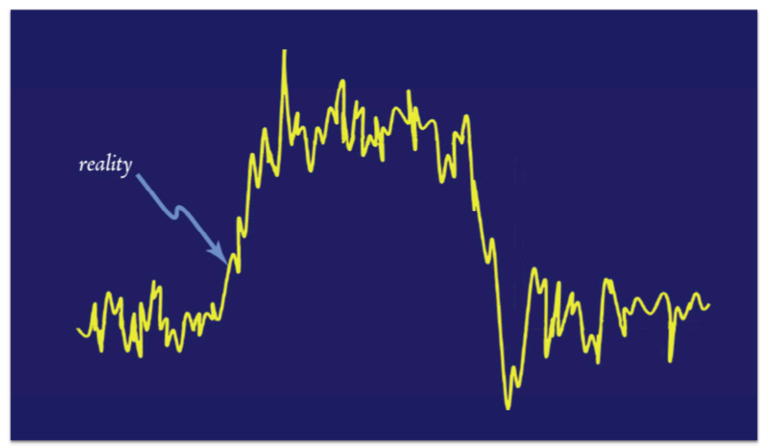


5. Haugeland doesn't tell you! (how to have digitality).
6. His account of digitality is a *very good* account of *what digitality gives you* — of **what digitality is for.**
7. Arguably (though I don't believe this), it is a *reasonable* account of
   — **What digitality is**
8. But it doesn't even *begin* to be an account of
   — **How digital systems can be constructed in the physical world!**

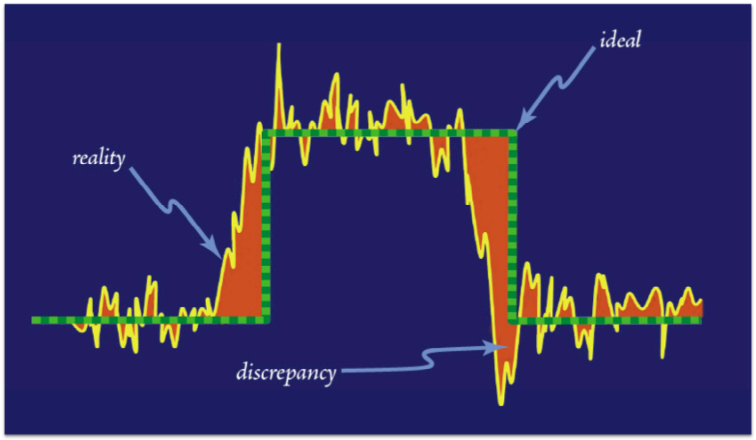
So how can we actually *have* digital systems—systems of such perfection?



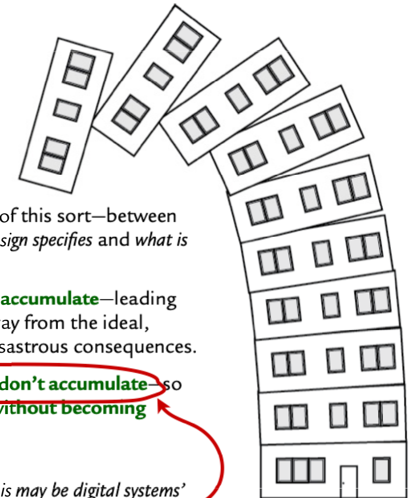
So how can we actually *have* digital systems—systems of such perfection?



So how can we actually *have* digital systems—systems of such perfection?



Discrepancy



1. All designed systems have **discrepancies** of this sort—between the ideal and reality—between *what the design specifies* and *what is actually the case* in the world.
2. Normally, these discrepancies **add up** or **accumulate**—leading the system to veer farther and farther away from the ideal, ultimately (if taken too far) leading to disastrous consequences.
3. But in digital systems **the discrepancies don't accumulate**—so they can be build of gazillions of parts, **without becoming unstable or losing their perfection!**
4. How do they do this?

This may be digital systems' most important property!

The fundamental strategy for achieving perfection

According to John Haugeland:

1. "**Digital**, like *accurate*, *economical*, or *heavy-duty*, is a mundane engineering notion, root and branch. It only makes sense as a practical means to cope with the vagaries and vicissitudes, the noise and drift, of earthly existence."
2. Positive "**read/write**" techniques
   a) '**Positive**': absolutely, totally, without qualification

Bad John! Just means affect and be affected by

✘ i. Cut a board 6′ long
✘ ii. Sort boards into 1′–2′, 2′–3′, 3′–4′?
✘ iii. Sort boards into a) 1.0′ ≤ length < 2.0′ b) 2.0′ ≤ length < 3.0′ c) 3.0′ ≤ length < 4.0′

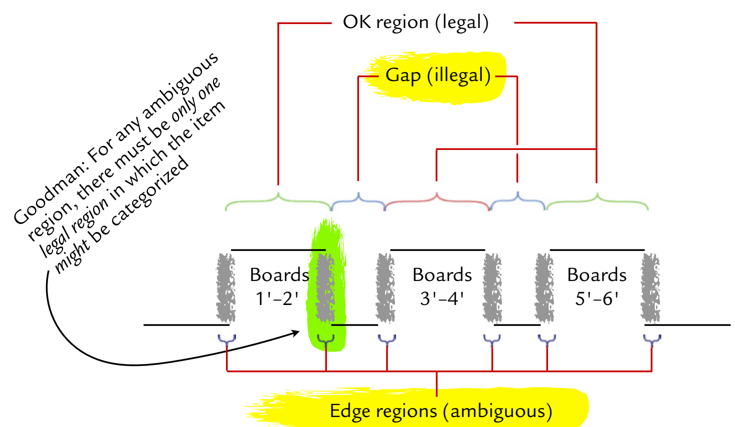
these are **practical impossibilities**

✔ iv. Sort boards into *(assuming there are no 2′–3′ or 4′–5′ boards)* a) 1.0′ ≤ length ≤ 2.0′ b) 3.0′ ≤ length ≤ 4.0′ c) 5.0′ ≤ length ≤ 6.0′

this works because there is a **gap**

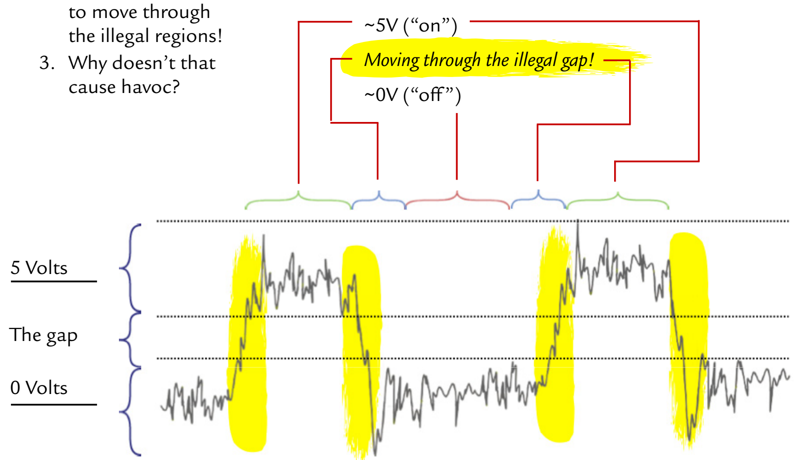
The critical gap

I.e., in a real-world digital system, there needs to be a *gap* separating the OK regions …

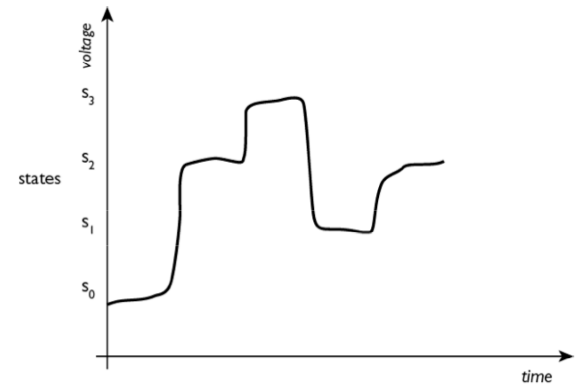


The gap (cont'd)

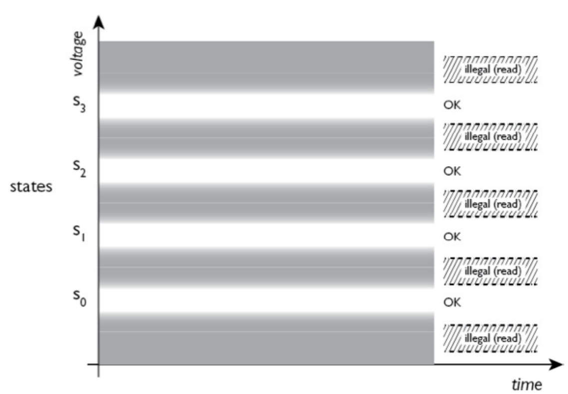
1. No matter what the encoding, there needs to be a *gap* separating the OK regions
2. But given that the underlying world is continuous, at some points the system has to move through the illegal regions!
3. Why doesn't that cause havoc?



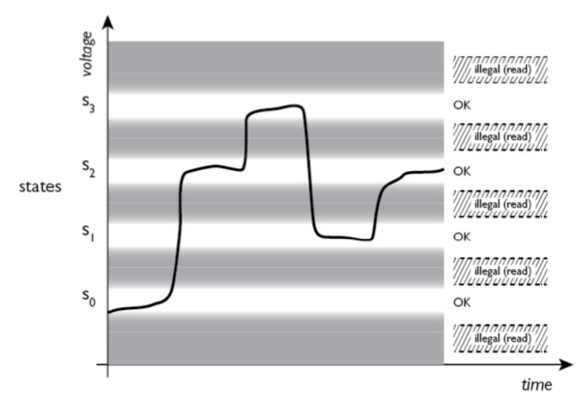
To build a computer, you need amplitude (voltage) and temporal gaps



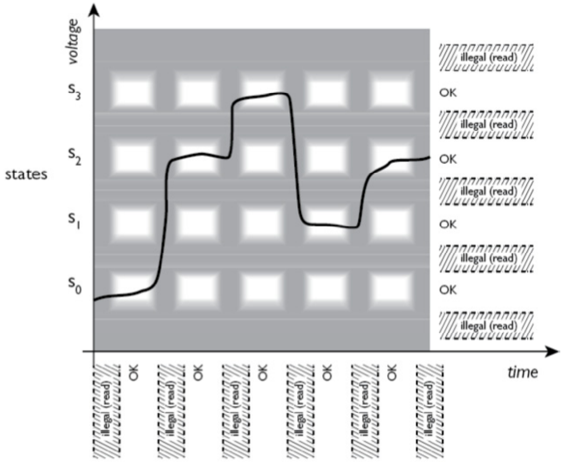
To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)



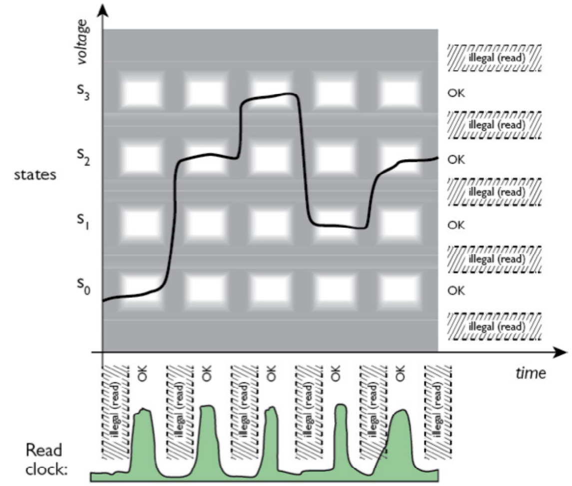
To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)



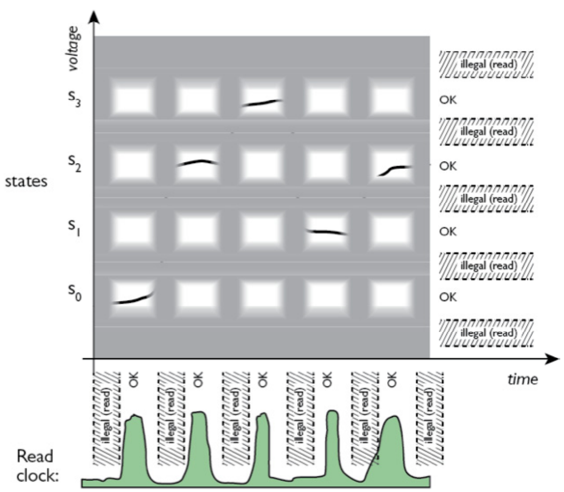
To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)



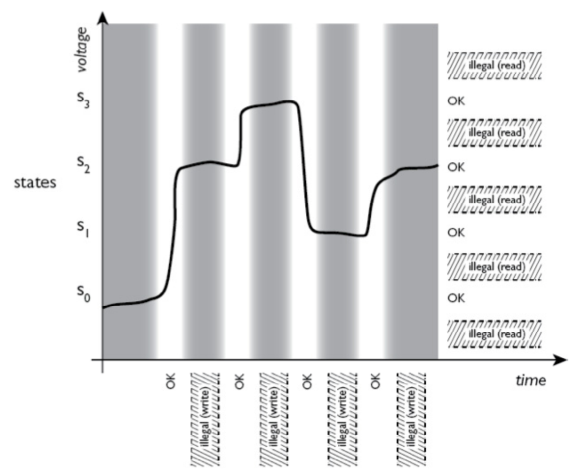
To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)



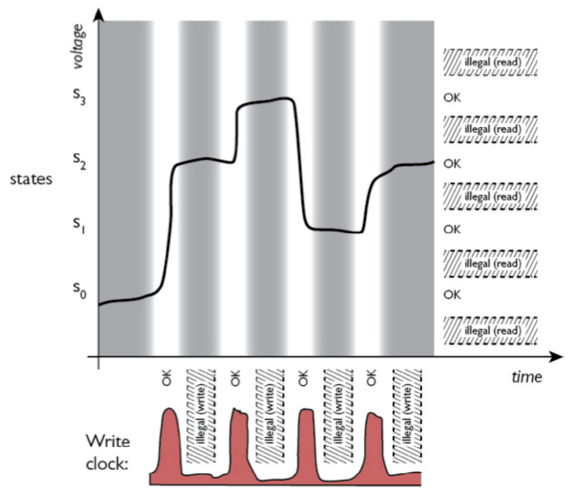
To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)



To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)

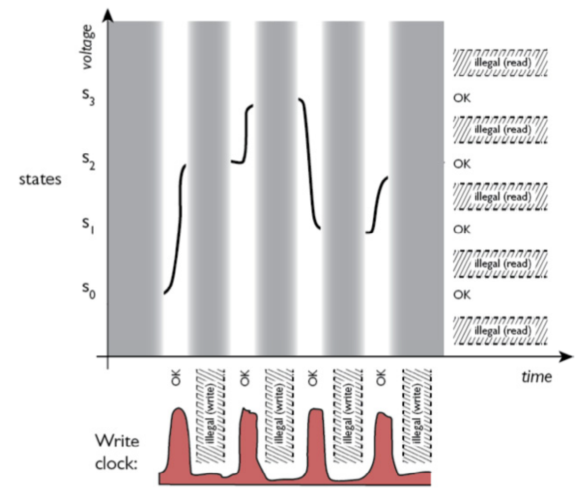


To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)



2019 · April · 17 Minds & Machines Lecture — B · 02

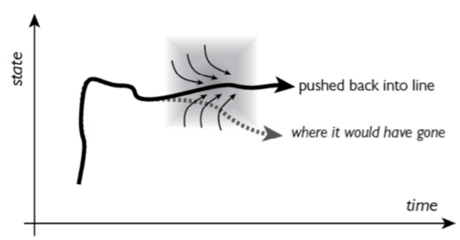
To build a computer, you need amplitude (voltage) and temporal gaps (cont'd)



2019 · April · 17 Minds & Machines Lecture — B · 02

Analysis

1. In sum, building a computer (dynamic digital system) requires **three references**:
   a. **Two cross-beating** (synchronized) **clocks**—for reading and writing, and
   b. A **multi-level amplitude (voltage) reference** (for the "values")
2. There is another essential ingredient, in order for a digital system to work: The values of the signal, at each step, have to be "beaten back" into the centre of the legal region.
3. This can only be done well with a **good, stable reference** for each "OK" value (it can be the same reference as the one used for reading, but it requires additional circuitry)



2019 · April · 17 Minds & Machines Lecture — B · 02

Analysis (cont'd)

4. Other engineering techniques are used, above the level of the bits, to ensure that the "perfect idealization" can be maintained.
5. Most important are various kinds of **redundancy**, which allow errors in individual bits—and even groups of bits—to be ignored and/or corrected (all physical media will lose bits in the end...)
6. Cf. original spec for the CD (from the President of Sony):
   a. It had to fit into a "DIN" radio space on the dashboard of American cars;
   b. It had to sound "as good" as the best analogue recording;
   c. It had to have enough room to encode a complete recording of Beethoven's 9th symphony; and
   d. It had to be able to recover (with no errors) from fingernail scratches across its surface (hundreds of bits wide!).

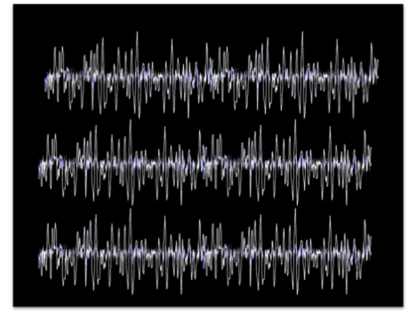
moth and rust do corrupt, in this world of ours

Analysis (cont'd)

7. In case all of these requirements are met, then Haugeland's criteria (of *perfect copyability*, *absolute determination of the types of tokens*, etc.) can be met.
8. In my opinion, this is **totally and utterly amazing**!
9. Imagine that King Arthur asked you to develop a system that met Haugeland's criteria of *absolute uniform and distinct types*, *absolute definite yes/no answers to questions of type and token*, *absolute reliability*, etc. — and to build it in such a way that not only could moderately complex systems be built (e.g., the structural beams in a skyscraper), but systems of 10^9, 10^{12}, 10^{15}, 10^{18} parts, would work reliably for 10^{17} or 10^{18} or more operations (a total complexity approaching ~10^{40}).
10. Wouldn't your immediate response be that in this imperfect world *such a thing cannot be done*?
11. Yet all contemporary computer systems meet this standard!

Postscript — The *dimensionality* of digitality

1. Systems are devices are *not*, in general
   a. **100% digital**, or
   b. **100% analog**
2. If general, the question of whether something is analog or digital has to be asked with respect to each of its constitutive dimensions.
3. So-called "*analog* TV," for example, is:
   a. **Analogue horizontally**, and
   b. **Digital vertically!**

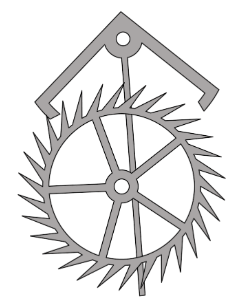


Postscript — The *dimensionality* of digitality (cont'd)

4. Similarly, only the second hand on the face of a traditional "analogue" watch is read in an analogue fashion; the other two hands are **read digitally**

Finally — some applications (and complexities) — cont'd

5. Perhaps even more surprisingly, the **mechanism** in traditional ("analog") mechanical clocks is in fact **digital!**
6. Their limitations (as accurate time-keepers) has to do with the escapement and balance spring, which constitute the "**temporal reference**"—which works by alternating between two discrete states.



Digital music ...

7. And so on...
8. If there were time, we could talk about CDs, and digital music, and why, at least from a theoretical point of view, no two pressings of the same CD (i.e. two physical tokens of a CD with "identical bitstreams") can produce exactly the same acoustic waveform ...
9. And other wonders of the digital world.

Perhaps later in the semester, if we have time ... but meantime ...

2019 · April · 17 Minds & Machines Lecture — B · 02

Return from our digression on ***digitality*** *(on Tuesday)*

GOFAI
(Good Old-Fashioned Artificial Intelligence)

Sept 28, 2017

1. As we have seen, GOFAI, the classical "architecture of mind" on which AI was founded, models the mind on a roughly "logic-based" conception
2. That model is in turn built on two (well-recognized) philosophical positions of **functionalism** and **multiple realisability**.
3. We'll start our analysis of GOFAI by looking at them.

Functionalism

1. As we've said, its focus on digital systems of representation, syntax, semantics, etc. (with their associated ability to deal with productivity, systematicity, and compositionality) is based on an interlocking set of 3 assumptions:
   a) We are not interested in mind *at the lowest level of physical realisation/implementation* (such as neuroscience)
   b) Nor are we interested in viewing the mind as something that can be comprehended *merely in terms of its external input/output relations in the world* (behaviourism)
   c) Rather, we should understand it at a **higher, more abstract** level, at which we can identify regularities between and among representational or "content-bearing" (semantically evaluable) states.
2. This leads to a recognition that we are interested in the mind as a mechanism at a **functional level of abstraction**
   a) Cf. radios and televisions ("power supplies," "amplifiers," etc., are *functionally individuated*)
   b) Cf. coke-machine change dispensers
   c) Cf. chairs, tables, and buildings?
3. Turing machines, computers, etc., are also functionally identified—built of components that are defined and classified according to *what they do.*

Multiple Realisability

1. Functional requirements can be met by many different physical substrates—a thesis known as **multiple realisability**
2. Intuitively, multiple realisability makes sense—you can build a Turing machine out of lots of different materials:
   a) Vacuum tubes
   b) Transistors
   c) Integrated circuits
   d) Tinker Toy (this was done at MIT)
3. Similarly for chess: you can play it with wooden pieces, on a computer, using helicopters (potentially with positions on the chess board separated by hundreds of miles), etc.
4. At a deeper level, though what exactly "multiple realisability" means isn't so clear:
   a) *Independent* of physical form?
   b) In *any* physical form?
   c) In *some* physical forms? (For example, it seems hard to implement a Turing machine out of nothing but gaseous oxygen.)
5. These are open questions: what the requirements are on **realisation**—or as computer scientists would say, on **implementation**—are not well-understood.

The Argument for Using Logic

We also saw that the GOFAI model is also based on a thesis that the best way to build a machine that can *reason intelligently about the world* is to *make it out of* **symbols that represent the world**.

In particular:

1. Because
   a. Intelligence involves being (able to be) *productive* and *systematic*, and
   b. The only idea anyone has ever had (including up to today!) about how to be productive and systematic is to be (or to be able to use language or representations that are) *compositional*,

   It is assumed that the representational symbol system must be a compositional one.
2. The most highly developed compositional knowledge representation language ever developed is (one or other variant of) **formal logic**.
3. The initial suggestion for constructing an intelligent machine, therefore, is to build it to use a **knowledge representation system** (or language) based on (compositional) logic.

Knowledge representation in GOFAI

1. Use logic, or a system closely based on logic, as the best knowledge representation language in which to represent the world, including all relevant entities and facts about the world relevant to the task at hand.
2. Take *kinship* as an elementary example
3. Some particular facts

Person(Pat)	— *Pat is a person*
Person (Hilary)	— *Hilary is a person*
∃x [Child(x, Hilary) ∧ Child(x, Pat)]	— *There is someone who is a child of Hilary and a child of Pat (i.e., Pat and Hilary have a child)*

4. Some additional facts we (as people) might *assume*, but that for computers need to be represented explicitly:

∀x [[∃y Child(x, y)] → Person(x)]	— *Every child is a person*
∀y [[∃x Child(x, y)] → Person(y)]	— *Only people have children*
∀x [∃y Child(x, y)]	— *Everyone is the child of someone*

In Passing — A (Quick) Glossary of Logical Syntax

Symbol	Meaning	Example	English
...(...)	Predication	Tall(Pat)	*Pat is tall*
...(..., ...)	Relations	Father(Llewellyn,Kat)	*Llewellyn is the father of Kat*
∀x [...]	"For all x ..."	∀x [Green(x)]	*Everything is green*
∃x [...]	"There exists x..."	∃y [Whale(y)]	*There is a whale*
∧	Conjunction	Young(x) ∧ Beautiful(x)	*x is young & beautiful*
∨	Disjunction	Even(z) ∨ Odd (z)	*z is even or z is odd*
¬	Negation	¬ Flies (Tweety)	*Tweety does not fly*
→ (or ⊃)	Implication	Child(x,y) → Parent(y,x)	*If x is y's child, then y is x's parent* or: "Child(x,y) *implies* Parent(y,x)
↔	"If and only if (iff)"	Sibling(x,y) ↔ Sibling(y,x)	*x is y's sibling iff y is x's sibling*
=	Identity	Cicero = Tully	*Cicero is the same as Tully*
≠	Non-identity	Vienna ≠ Venice	*Vienna is not Venice*
⊢	Derivation	$S_1, S_2, S_3 \vdash S$	S *can be (formally!) derived from* S_1, S_2, *&* S_3
⊨	Entailment	$S_1, S_2, S_3 \models S$	S_1, S_2, *&* S_3 *(semantically!) entail* S

Knowledge Representation in GOFAI (cont'd)

5. And perhaps a whole lot more additional facts as well:

∀x,y [Child(x,y) → Parent(y,x)]	— *The inverse of 'child' is 'parent'*
∀x [Person(x) → [Male (x) ∨ Female(x)]	— *Every person is either male or female*
∀x ¬ [Male(x) ∧ Female (x)]	— *No one is both male and female*
∀x, y [[Parent(x,y) ∧ Female(x)] ↔ Mother(x,y)]	— *A mother is a female parent*
∀x, y [[Parent(x,y) ∧ Male(x)] ↔ Father(x,y)]	— *A father is a male parent*
∀x, y, z [[Child(x,z) ∧ Child(y,z) ∧ x≠y] ↔ Sibling(x,y)]	— *(Different!) children of the same parent are siblings*
∀x, y [Sibling(x,y) → Sibling(y,x)]	— *Sibling is symmetrical*
∀x, y [[Sibling(x,y) ∧ Male(x)] ↔ Brother(x,y)]	— *Male siblings are brothers*
∀x, y [[Sibling(x,y) ∧ Female(x)] ↔ Sister(x,y)]	— *Female siblings are sisters*
∀x, y, z [[Child(x,y) ∧ Brother(z,y)] ↔ Uncle(z,x)]	— *Uncles are brothers of parents*
∀x, y, z [[Child(x,y) ∧ Sister(z,y)] ↔ Aunt(z,x)]	— *Aunts are sisters of parents*

... and so on!

6. It takes an **enormous number of logical statements** (or knowledge representation expressions) to represent even the *simplest* domains!

— Cf. my 1990 "The Owl and the Electric Encyclopedia" (on Blackboard)

Inference and Search

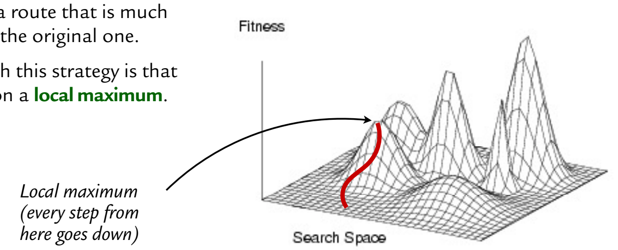
1. It is not enough merely to *represent the knowledge* that a GOFAI system needs.
2. You also have to give it instructions or algorithms so that it will **do something**.
3. Logical inference rules will allow you to draw sound conclusions—but by themselves they don't tell you which rules to apply. Without some *strategy*, a random inference regimen might draw endless irrelevant conclusions.
4. Rather, what is needed is some way to specify what you are looking for, or what you want to figure out (whether something is true or not, e.g.)
5. Newell and Simon formulated this issue in terms of **search**.
6. A **problem space**, they said, was a **search space**.
7. Solving a problem involved searching the space of possible solutions to find one that worked.

Search — some simple examples

1. One (not very good) strategy is to use **blind search**: just search through the possible solutions, in some order, and try each one.
   - E.g., the problem is to find a store in Toronto that sells madeleines
   - Blind search: start at some random place; go to a random store on that street, and ask the people there whether they sell madeleines. If they don't, pick another random store in Toronto (that you haven't already visited) and go there. If there is any store in Toronto that *does* sell madeleines, you are guaranteed to find it…*eventually!*
2. Another strategy, if you have a way to measure or know whether a given step takes you closer to the goal, is to make a step in that direction, and then repeat.
   - Example: if you want to get to the highest point in Ontario, start where you are, look around you, and step in a direction that moves you upwards (at least a bit). It is basically a strategy of *iterative improvement.*
   - More generally, the strategy is to start with some solution (perhaps a random one), and then attempt to find a better solution by incrementally changing your proposed solution in a simple (usually local) way. If the change produces a better solution, an incremental change is made to the new solution, repeating until no further improvements can be found.
   - This is called **hill-climbing**.

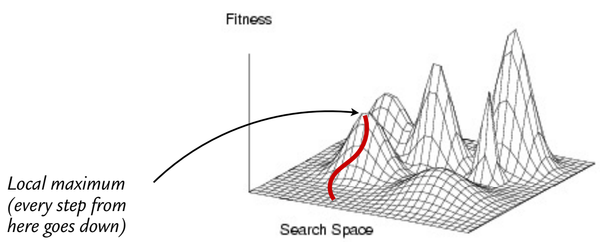
Search — some simple examples (cont'd)

3. Hill climbing can be used, for example, to find the *shortest route* that visits all of a specified set of destinations (known as the "**traveling salesman problem**").
   a) Start by finding *any* route that visits all of the destinations—and measure its length.
   b) Then adjust the route a bit (e.g., change the order of two of the destinations).
   c) If that is shorter, take that and repeat the process of adjustment. If it is not shorter, then abandon it, and make some other incremental change.
   d) Eventually, the idea is, you will end up with a route that is much shorter than the original one.
4. The problem with this strategy is that it will get stuck on a **local maximum**.



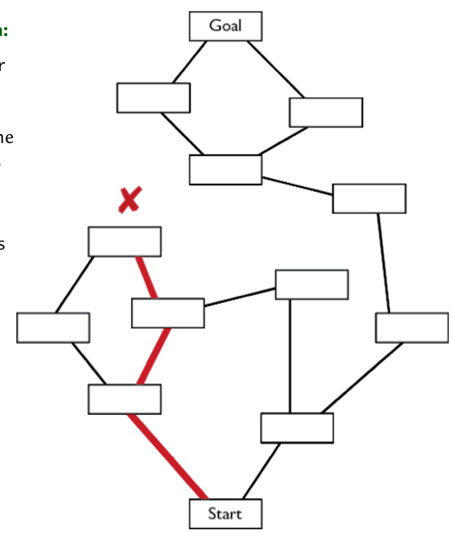
Search — some simple examples (cont'd)

5. One strategy to deal with local maxima (and other forms of failure en route) is to **back up**, when they are encountered, and to try another path towards the goal.
6. This, too, can be horribly inefficient, but in the end, if a solution exists, it should be found.
7. Often—as for example in this hill-climbing case—it can be challenging to know *how far to back up*, to that you don't keep climbing the same local hill. But sometimes backing up to the "last decision point" is easier…



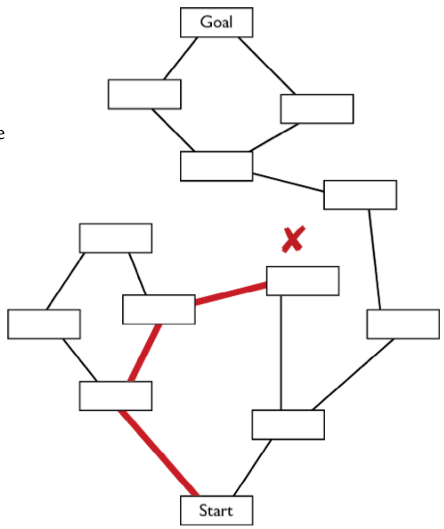
Forward Algorithm:

1. If there is one or more steps that move towards the goal, take the step that moves closest to the goal.
2. If no step moves closer to the goal, mark this route as bad; back up to the last decision point; and repeat.
3. If goal is reached, stop (success!).



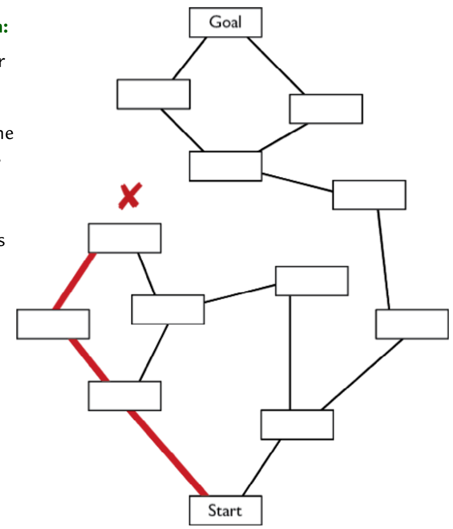
Forward Algorithm:

1. If there is one or more steps that move towards the goal, take the step that moves closest to the goal.
2. If no step moves closer to the goal, mark this route as bad; back up to the last decision point; and repeat.
3. If goal is reached, stop (success!).



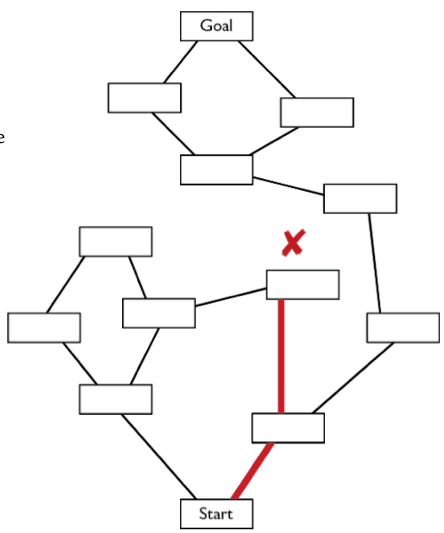
Forward Algorithm:

1. If there is one or more steps that move towards the goal, take the step that moves closest to the goal.
2. If no step moves closer to the goal, mark this route as bad; back up to the last decision point; and repeat.
3. If goal is reached, stop (success!).



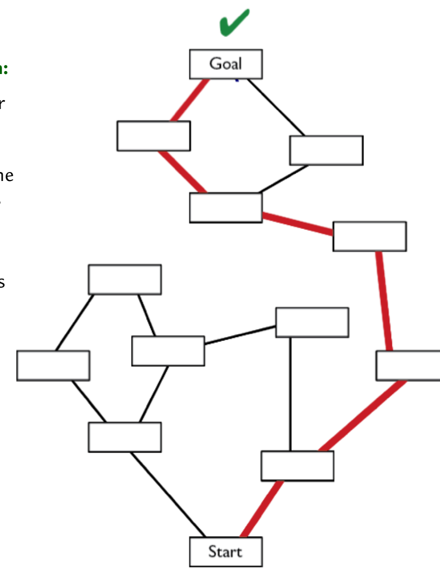
Forward Algorithm:

1. If there is one or more steps that move towards the goal, take the step that moves closest to the goal.
2. If no step moves closer to the goal, mark this route as bad; back up to the last decision point; and repeat.
3. If goal is reached, stop (success!).



Forward Algorithm:

1. If there is one or more steps that move towards the goal, take the step that moves closest to the goal.
2. If no step moves closer to the goal, mark this route as bad; back up to the last decision point; and repeat.
3. If goal is reached, stop (success!).

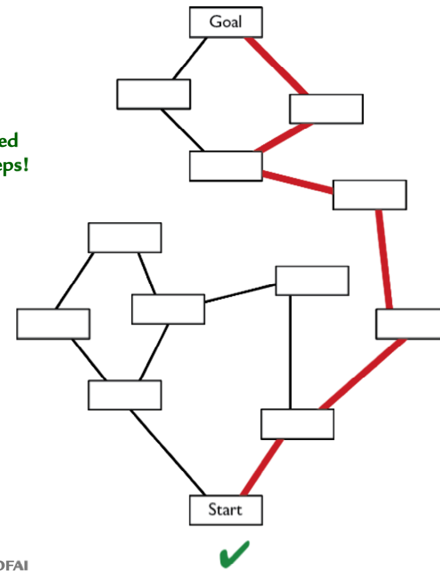


Success reached after a total of 13 steps!

Search — some simple examples (cont'd)

7. In some cases, it may be more efficient to work **backwards** from the goal, rather than forward from the starting point.

Success reached after just 6 steps!



Backward Algorithm:

1. Start at the goal
2. If there is one or more steps that moves towards the start, take the step that moves closest to the start.
3. If no step moves closer to the start, mark this route as bad; back up to the last decision point; and repeat.
4. If start is reached, stop (success!).

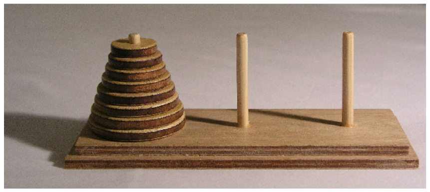
General Problem Solver (GPS)

1. These and similar search strategies were used in a program written in 1959 by Herbert Simon, Cliff Shaw, and Alan Newell, called the "General Problem Solver" (GPS).
2. Whereas previous software had been written with very specific goals, and to solve very specific problems, GPS (at the name implies) was intended to solve nearly any problem.
3. GPS certainly couldn't solve any problem, but it was able to do the following sorts of things:
   a) Arithmetic "word problems", such as:
      — *Rhonda has 12 marbles more than Douglas. Douglas has 6 marbles more than Bertha. Rhonda has twice as many marbles as Bertha has. How many marbles does Douglas have?*
      — *Brothers and sisters have I none. But that man's father is my father's son. Who is that man?*

      This problem is easily solved nowadays, e.g., by Wolfram Alpha
   b) Simple "means-ends analysis" (e.g., their example of taking a son to school)

      See this link.
   c) The "Tower of Hanoi"

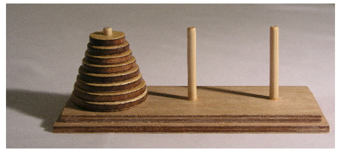
Tower of Hanoi (or Tower of Brahma)



There are three vertical rods, and a number of disks of different sizes, which can slide onto any rod. At the outset, all the disks are on the left rod, in ascending order of size, with the smallest at the top. The objective is to move the entire stack to the right hand rod, obeying these rules:

1. Only one disk can be moved at a time
2. Each move consists of taking the uppermost disk on one rod and placing it on another rod, on top of any disks already on that rod.
3. No disk may be placed on top of a smaller disk.

In general, the problem can be solved in $2^n - 1$ moves, where *n* is the number of disks.



Tower of Hanoi (Brahma) — cont'd

The puzzle was invented in 1883 by French mathematician Édouard Lucas. It is associated with a mythical legend,* that runs as follows (though there are many versions):

> "*In the great temple at Benares there is a brass plate in which are fixed three diamond needles. At the creation of the world God placed sixty-four discs of pure gold on one of these needles, arranged in order of size, the largest disc resting on the brass plate. Day and night the priests move the discs, one at a time, from one diamond needle to another according to the fixed and immutable laws of Brahma, never placing a larger disk on a smaller one. When all sixty-four discs have been transferred from the needle on which God first placed them to one of the other needles, then tower, temple, and Brahmins alike will crumble into dust, and with a thunderclap the world will vanish.*"[1]

If the legend were true, and if the priests were able to move disks at a rate of one per second, using the smallest number of moves, it would take them $2^{64}-1$ seconds, or roughly 585 billion years—or 18,446,744,073,709,551,615 turns to finish, or about 127 times the current age of the sun, or 42 times the age of the universe.†

**As Wikipedia says: "It is not clear whether Lucas invented this legend or was inspired by it"*

†Even coded maximally efficiently on a fast contemporary Intel processor, it would take about 6,000 years

Some issues with logical representation

1. All assumptions have to be **represented explicitly** (in contrast to people, who *seem* to be able to assume without reasoning). E.g., subcategories:

$$\forall x\ [\ \mathrm{Colour}(x) \rightarrow [\ \mathrm{Red}(x) \lor \mathrm{Orange}(x) \lor \mathrm{Yellow}(x) \lor \mathrm{Green}(x) \lor \mathrm{Blue}(x) \lor \mathrm{Violet}(x)\ ]$$

2. But this allows something to be more than one colour at once:

$$\exists x\ [\ \mathrm{Orange}(x) \land \mathrm{Green}(x) \land \mathrm{Violet}(x)\ ]\ ✘$$

3. To block this, we have to **state explicitly** that something can't be more than one colour:

$$\begin{array}{ll}
\forall x\ [\ \neg[\ \mathrm{Red}(x) \land \mathrm{Orange}(x)\ ] & \land \neg[\ \mathrm{Orange}(x) \land \mathrm{Violet}(x)\ ] \land \\
\quad \neg[\ \mathrm{Red}(x) \land \mathrm{Yellow}(x)\ ] & \land \neg[\ \mathrm{Yellow}(x) \land \mathrm{Green}(x)\ ] \land \\
\quad \neg[\ \mathrm{Red}(x) \land \mathrm{Green}(x)\ ] & \land \neg[\ \mathrm{Yellow}(x) \land \mathrm{Blue}(x)\ ] \land \\
\quad \neg[\ \mathrm{Red}(x) \land \mathrm{Blue}(x)\ ] & \land \neg[\ \mathrm{Yellow}(x) \land \mathrm{Violet}(x)\ ] \land \\
\quad \neg[\ \mathrm{Red}(x) \land \mathrm{Violet}(x)\ ] & \land \neg[\ \mathrm{Green}(x) \land \mathrm{Blue}(x)\ ] \land \\
\quad \neg[\ \mathrm{Orange}(x) \land \mathrm{Yellow}(x)\ ] & \land \neg[\ \mathrm{Green}(x) \land \mathrm{Violet}(x)\ ] \land \\
\quad \neg[\ \mathrm{Orange}(x) \land \mathrm{Green}(x)\ ] & \land \neg[\ \mathrm{Blue}(x) \land \mathrm{Violet}(x)\ ] \land \\
\quad \neg[\ \mathrm{Orange}(x) \land \mathrm{Blue}(x)\ ] &
\end{array}$$

Some issues with logical representation (cont'd)

4. In general, such issues of **non-identity** are a pain if they have to be represented explicitly. E.g., suppose we want to represent that there are *three* things:

$$\exists x_1 \land \exists x_2 \land \exists x_3 \land [\neg[\ x_1 = x_2\ ] \land \neg[\ x_1 = x_3\ ] \land \neg[\ x_2 = x_3\ ]]$$

5. That might be OK—though it is pretty awkward.
6. But now suppose we want to represent that there are *eleven* things:

$$\begin{array}{l}
\exists x_1 \land \exists x_2 \land \exists x_3 \land \exists x_4 \land \exists x_5 \land \exists x_6 \land \exists x_7 \land \exists x_8 \land \exists x_9 \land \exists x_{10} \land \exists x_{11} \land \\
\neg[[\ x_1 = x_2\ ] \lor [\ x_1 = x_3\ ] \lor [\ x_1 = x_4\ ] \lor [\ x_1 = x_5\ ] \lor [\ x_1 = x_6\ ] \lor [\ x_1 = x_7\ ] \lor [\ x_1 = x_8\ ] \lor \\
\quad [\ x_1 = x_9\ ] \lor [\ x_1 = x_{10}\ ] \lor [\ x_1 = x_{11}\ ] \lor [\ x_2 = x_3\ ] \lor [\ x_2 = x_4\ ] \lor [\ x_2 = x_5\ ] \lor [\ x_2 = x_6\ ] \lor \\
\quad [\ x_2 = x_7\ ] \lor [\ x_2 = x_8\ ] \lor [\ x_2 = x_9\ ] \lor [\ x_2 = x_{10}\ ] \lor [\ x_2 = x_{11}\ ] \lor [\ x_3 = x_4\ ] \lor [\ x_3 = x_5\ ] \lor \\
\quad [\ x_3 = x_6\ ] \lor [\ x_3 = x_7\ ] \lor [\ x_3 = x_8\ ] \lor [\ x_3 = x_9\ ] \lor [\ x_3 = x_{10}\ ] \lor [\ x_3 = x_{11}\ ] \lor [\ x_4 = x_5\ ] \lor \\
\quad [\ x_4 = x_6\ ] \lor [\ x_5 = x_7\ ] \lor [\ x_4 = x_8\ ] \lor [\ x_4 = x_9\ ] \lor x_4 = x_{10}\ ] \lor [\ x_4 = x_{11}\ ] \lor [\ x_5 = x_6\ ] \lor \\
\quad [\ x_5 = x_7\ ] \lor [\ x_5 = x_8\ ] \lor [\ x_5 = x_9\ ] \lor [\ x_5 = x_{10}\ ] \lor [\ x_5 = x_{11}\ ] \lor [\ x_6 = x_7\ ] \lor [\ x_6 = x_8\ ] \lor \\
\quad [\ x_6 = x_9\ ] \lor [\ x_6 = x_{10}\ ] \lor [\ x_6 = x_{11}\ ] \lor [\ x_7 = x_8\ ] \lor [\ x_7 = x_9\ ] \lor [\ x_7 = x_{10}\ ] \lor [\ x_7 = x_{11}\ ] \lor \\
\quad [\ x_8 = x_9\ ] \lor [\ x_8 = x_{10}\ ] \lor [\ x_8 = x_{11}\ ] \lor [\ x_9 = x_{10}\ ] \lor [\ x_9 = x_{11}\ ] \lor [\ x_{10} = x_{11}\ ]]
\end{array}$$

7. There must be a better way ;-)

Some issues with logical representation (cont'd)

7. One possibility—use sets, plus some simple arithmetic operations:

$$\exists s \land \mathrm{Set}(s) \land [[\ \mathrm{Cardinality}(s) = 11]$$

8. As normally axiomatized, sets assume the principle that there are no duplicates, but the fact can be represented explicitly (if awkwardly):

$$\begin{array}{l}
\forall x\ \forall s\ [[\mathrm{Set}(s) \land [\ x \in s\ ]] \rightarrow \\
\qquad [\ \forall t\ \ [\mathrm{Set}(t) \rightarrow \\
\qquad\qquad [[\ \forall y\ [\ y \in t\ ] \rightarrow [[\ y \in s\ ] \land \neg[\ y = x\ ]]] \land \\
\qquad\qquad\ [\ \forall y\ [\ y \in s\ ] \rightarrow [\neg[\ y = x\ ] \rightarrow [\ y \in t\ ]]]] \rightarrow \\
\qquad\qquad\qquad [\ \mathrm{Cardinality}(s) = [\ 1 + \mathrm{Cardinality}(t)\ ]]]]]
\end{array}$$

9. Another non-identity that we need to represent: that different constants (differently named identifiers) represent different people. At the moment, our system would license:

$$\mathrm{Person}(\mathrm{Hilary}) \land \mathrm{Person}(\mathrm{Pat})$$

10. Can the system assume that Hilary ≠ Pat, because they have different names?

Some issues with logical representation (cont'd)

11. No! The problem is that the **name of the person** is not automatically assumed to be the **name of the constant** (in logic) **that designates them**. The normal thing that one needs to do is to represent names (yes) explicitly:

    Name(Hilary, "Hilary") ∧ Name(Pat, "Pat")

12. On such a proposal, we would then need to *represent the uniqueness of names*:

    $\forall x_1\ \forall x_2\ \forall n_1\ \forall n_2\ [[[\ \text{Name}(x_1, n_1) \wedge \text{Name}(x_2, n_2)] \wedge [\ n_1 \neq n_2\ ]] \rightarrow [\ x_1 \neq x_2\ ]]$

13. Having two different "name spaces" isn't very convenient, though.
14. These sorts of consideration lead people to *add facilities* to logic, such as the ability to *refer to the names*, or to *refer to properties as such*, etc.
15. Instead of representing **transitivity** explicitly, for example, such as in:

    $\forall x\ \forall y\ \forall z\ [[\ \text{Near}(x,y)\ \text{Near}\ (y,z)] \rightarrow \text{Near}(x,z)\ ]$

    if we had that kind of "higher-order" access, we could say something like this (assuming that the inference system was able to handle this appropriately):

    Transitive(↑Near)

2017 · Sep · 28 PHL342 · Minds & Machines Lecture · 04 (b)

Some issues with logical representation (cont'd)

16. This approach would also allow us to deal with the subcategories (assuming that we represented 'ExclusiveSubcategories' appropriately):

    ExclusiveSubcategories(↑Colour, ↑Red, ↑Orange, ↑Yellow, ↑Green, ↑Blue, ↑Violet)

17. Other sorts of operators that are added to complicate logics:

    ◇P — for **possibly** P (i.e., P is **possibly true**) — i.e., ◇(Hillary, President)
    □P — for **necessarily** P (i.e., P is **necessarily true**) — i.e., □(4 = +(2, 2))

18. You can imagine other sorts of operators that would (at least seem to be) useful in modelling human cognition: **generally** or **usually P**, **officially P,** etc. (you won't often see these talked about in discussions of logic, but it is not clear that a plausible model of human cognition can avoid them...)

2017 · Sep · 28 PHL342 · Minds & Machines Lecture · 04 (b)

Intensionality and Opaque Contexts

19. Another huge complex topic has to do with the fact that various operators, such as *belief* and *know*, are—as it is said—**opaque** or **non-truth-functional**
20. Thus suppose that the following is true (since Tully is Cicero)

    Smart(Cicero) ↔ Smart(Tully)

    This does not imply that the following is true:

    Believes(Randy, Smart(Cicero)) ↔ Believes(Randy, Smart(Tully))

    Rather, the following might be true:

    Believes(Randy, Smart(Cicero)) ∧ ¬Believes(Randy, Smart(Tully))

21. The second argument position of Believes(x,P) is called an **intensional context**, since it is opaque or non-truth-functional:

    Believes(x, __________)

    because the truth-value of the whole sentence can change, depending not only on the *truth* of the embedded sentence, but on its **intension**—on what it **means**.

    — We haven't yet talked in this class about what "meaning" is; perhaps we will get to that (though it is a deucedly tricky subject!)

2017 · Sep · 28 PHL342 · Minds & Machines Lecture · 04 (b)

Defaults and Non-Monotonic Reasoning

22. Yet another issue: how to deal with what is called **default reasoning**—the ability to use a generalization or implication that is usually true, or that can *pretty much be assumed to be true unless you explicitly know something to make you think otherwise*, etc.

    Bird(x) → Flies(x)

    This is not absolutely true, of course, since x might be a penguin or ostrich.

23. As we have done in previous cases, we could try to put all the blocking conditions in explicitly (the way that banks and insurances companies do...):

    Bird(x) → [¬[Penguin(x) ∨ Ostrich(x)] → Flies(x)]

24. But in practice this has proved to be extremely awkward—often impossible, in fact, in part because people often do not *know* all the conditions. And suppose that one encounters a *particular* bird Tweety, who cannot fly (for who knows what reason). Does one want to revise one's *general* statement about birds flying to have to mentioned Tweety?

    Bird(x) → [¬[Penguin(x) ∨ Ostrich(x) ∨ [x = Tweety]] → Flies(x)]

    Surely not!

Defaults and Non-Monotonic Reasoning (cont'd)

25. Such issues have led people to introduce and explore logic-based reasoning systems that can handle such defaults, which typically have to be marked *as* defaults, to make things work—something like:

    ■ [Bird(x) → Flies(x)]

26. The idea would be that in the *absence of over-riding information*, the following inference would be licensed:

    Bird(Daffy) ⊢ Flies(Daffy)

27. On the other hand, suppose that we had these facts represented:

    ∀x [Penguin(x) → ¬Flies(x)]

    ∀x [Ostrich(x) → ¬Flies(x)]

28. Then if we know (i.e., represent) that Oswald is an Ostrich, then the inference that Oswald can fly would be blocked:

    Ostrich(Oswald) ⊬ Flies(Oswald)

29. Instead, we could conclude the opposite:

    Ostrich(Oswald) ⊢ ¬Flies(Oswald)

The Frame Problem

32. Yet another issue, tied to the fact that we are talking about live, temporal inference (rather than the detached proving of theorems) has to do with a system's response to *change*.

33. The most famous formulation of this problem is called the **frame problem**, having to do (to put it rather generally) with how to determine, if a change happens *what beliefs are thereby impacted*.

34. For example, if Randy goes to Vancouver, then (a) in the following will presumably *not* change, (b) *will* change, and (c) *might* change (though figuring out whether it will or not may depend on insight and arbitrarily complex reasoning).

    a. Person(Randy)

    b. Coming-to-dinner-tomorrow(Randy)

    c. Get-along-OK(Bobbie, Frankie)

The Frame Problem (cont'd)

35. Attempting to develop formal mechanisms for coping with the frame problem has been a *huge industry* in (GOFAI-oriented) AI. Among other things, it brings forward the issue of **relevance** — of when one belief (α) is relevant to another one (β).

36. While a number of (devilishly clever) mechanisms have been proposed, one can say in general that the problem of relevance—and the frame problem—have not been "solved" in any very thorough way.

37. More seriously, there is no consensus on whether they will *ever* be solved in a system that is based (like GOFAI) on explicit representation. (We will get back to this more next week when we talk about critiques of GOFAI).

Some issues with logical representation (cont'd)

... and so on!

1. These and myriad other issues of logical representation have to be dealt with in all GOFAI knowledge representation systems.
2. And yet ... it is not clear that any of the other "architectures of mind" that we will examine in part III of the course have dealt with any, let alone all, of the challenges we have just rehearsed.
3. Even the most embodied, situated, extended, enactive, emergent, conscious, deep-learning based synthetic creature will have to deal with all of the things that GOFAI ran up against.
4. What was damaging for GOFAI was the sense that it was *not addressing these issues in a compelling or generalizable way*. It was *how*, not *that*, it faced them. We deal with all of these things, and any serious "AI" will need to as well...
5. The question is whether there is a *better* (not handwaving!) way to approach them, that will work better than GOFAI. We will talk about this more in Part IV of the course—but there is no evidence that anyone has yet articulated a better strategy.
   — And note that knowledge representation is rising in importance once again—e.g., in Google's Knowledge Graphs...
6. The fundamental moral that should be taken from all of these examples is one of **humility** in the face of the dauntingly impressive things that our minds—which is to say, we—are able to do so flexibly, transparently, and almost instantaneously.

Have a good Weekend!

GOFAI Summary

(plus conversations among colourful AI personalities)

1. As we've said, there are
   a) Things you need to *understand* in this class (everything!), and
   b) Things you explicitly need to *remember or know*, based on this understanding.
2. On Thursday, we covered a lot of details about GOFAI and its reliance on logic and knowledge representation.
3. We'll start today going over the morals we should take from GOFAI—closer to the things about GOFAI that you should *know*.

Five morals from GOFAI's logic-based knowledge representation

Moral #1: Epistemic subtlety

a. The range of issues that knowledge deals with is **extraordinarily impressive**:
   i. Identity and non-identity
   ii. Quantification
   iii. Logical operators (and, not/negation, implies, etc.)
   iv. Sets
   v. Opacity (and intensional contexts)
   vi. Categories and subcategories
   vii. Possibility and necessity
   viii. Default reasoning
   ix. Relevance and the frame problem
   x. … etc.

b. If—as logicists and GOFAI adherents argue—these are all part and parcel of human thinking, then **any plausible cognitive architecture** will have to deal with them.

Five morals from GOFAI's logic-based knowledge representation (cont'd)

Moral #2: What someone knows vs. how someone thinks

a. As we saw when we talked about logic (Lecture 03a, Sept 21), logic is a normative enterprise, with norms applying to both *expressions* (sentences) and *forms of inference*.

b. Both norms are defined in terms of **truth:**
   i. Sentences should be *true*;
   ii. Inference should show what else is true (or false), depending on whether the sentences are true (or false). That is: it should be *truth-preserving.*

c. However logic was never designed to deal with **practical reasoning**.
   i. Suppose one believes "P" and that "$P \supset \neg Q$"
   ii. E.g., "Kim is paralyzed" and "if Kim is paralyzed then Kim can't run"
   iii. Does that mean that one should believe "$\neg Q$"? (e.g, that "Kim can't run")
   iv. *Not necessarily!* All logic is telling you is that "P" and "Q" are **incompatible**.
   v. One may have very good reasons to believe Q (suppose you *saw* Kim run!)
   vi. All you can conclude from logic is that you *should not believe all three of "P", "Q", and "$P \supset \neg Q$" at once!* One of them must go!

Five morals from GOFAI's logic-based knowledge representation (cont'd)

Moral #2: What someone knows vs. how someone thinks (cont'd)

d. *Which sentence you should discard, in case of a contradiction* (which of "P", "Q", and "$P \supset \neg Q$", in the example) is an independent issue. (In the example, perhaps you should conclude that Kim is not paralyzed after all.)

e. Logic can't identify what's right and what's wrong; nor can it say what to believe.

f. Rather, logic is a theory of the **relations among truth-evaluable sentences**

g. But an theory of mind must be a theory of truth *and of thinking*

h. The requirements for *how to think* go well beyond those illuminated by logic.

Five morals from GOFAI's logic-based knowledge representation (cont'd)

Moral #3: Relentless explicitness

a. Logic is also *relentlessly explicit*—requiring that everything of any conceivable salience be explicit coded in formulae.

b. This is explicitness of **ingredients** within the machine (or person); it does not mean that the (contents of the) representations are *explicit for the system* (machine, person) thereby constituted.

c. There are c. 100,000,000,000 neurons in each person (10^{11}), and 1,000 times that many connections between and among them (10^{14}).

d. Perhaps we (minds) are based on *massive numbers of explicit encodings*.

e. It must be admitted, though, that such a huge amount of explicitness seems at odds with our sense of what constitutes "reasonable" reconstruction of human mental capacities of which we are consciously aware.

This is a theme we will see several more times in this course: relations between:

i) *What is true of our inner workings ("subpersonally")*

vs.

ii) *What is phenomenologically and consciously apparent to us as whole persons*

Five morals from GOFAI's logic-based knowledge representation (cont'd)

Moral #4: Exactness

a. There is also something "absolute" (binary) about logical encodings: things either *are or are not true.* For any x: either P(x) or ~P(x)

b. There doesn't seem to be any room for *gradualness*, *vagueness*, things being *"more or less" so*, etc.

c. The exactness (sharp boundaries) are not just in the *representations* (formulae, sentences). Logic—and logic-based knowledge representation systems—also seems to assume that the **world itself also comes with sharp boundaries**, between and among *things, properties, relations, etc.*

d. Most semantical accounts of logic are also based on a very definite and exact **ontological** picture of the world: discrete objects, properties, relations—and perhaps sets, facts (or propositions), etc.

e. When we turn to critiques, we will see that the definiteness of this classical ontological picture has come under attack from a wide variety of sources.

f. How to represent and reason about a less definitely "carved up" world is a major challenge facing any proposed representational scheme.

Five morals from GOFAI's logic-based knowledge representation (cont'd)

Moral #5: Organization and Complexity

a. Another issue that comes up, in using logic-based representation languages, is the issue of **managing huge amounts of data**.

b. Logic, per se, does not deal with considerations of how knowledge is **organized**, how one finds sentence or formulae to use in inferences, etc.

c. In its use in logic, mathematics, and philosophy, logic-based systems are typically *very small*—a few or perhaps 100 representations.

   i) Cf. modal logic S5, which consists of just a handful of axioms

d. Present-day computer systems are **vastly larger**. Consider that the Android operating system (running on phones) is about *12 million lines of code*, and Windows, closer to *50 million*.

e. What it means to represent a vastly complex world in a vastly complex representation system remains an open question…not something to which logic-based systems suggest any ready answers.

You can't really claim to be a programmer unless you have written a program at least several thousand lines long!

Summary

1. A lot more can be said about the epistemic, inferential, ontological, and semantic properties of logic. We will return to some of them when we critique the entire GOFAI framework.
2. But these examples should convey a sense of the issues.
3. While you may well agree with the critiques… keep in mind its power as well.

In passing—a comment on semantics…

1. If we are building computer systems **for us to use**—i.e., systems to help us humans out—then it is fine for the semantic interpretation of the system's symbols and behaviour (and hence its overall intelligibility) to be tacitly attributed by us.
2. We don't actually care whether *Siri* knows what a pizza is, or "the nearest bathroom."
   a) How it arises, and how it is established
   b) That is, for systems we use, it is fine for the semantics to be "derivative"
3. For the study of the **mind**, however, and as regards the prospects of constructing a genuinely (i.e., *authentically*) intelligent device, questions about the origin and sustenance of the semantic interpretation do have to be answered:
   a) How it arises, and how it is established
   b) How it can act as an overarching normative constraint on the system's design and behaviour
   c) How the causal mechanisms can ensured to remain true to it

Some Conversations with Early AI Personalities
Eliza, Parry, and Racter

1. These examples taken from Güven Güzeldere and Stefano Franchi, "Dialogues with Colorful Personalities of Early AI," in an issue of the *Stanford Electronic Humanities Review* they edited, entitled "Constructions of the Mind," volume 4, issue 2.
   — available at: http://www.stanford.edu/group/SHR/4-2/text/toc.html
2. These examples are dated, awkward, and rather madcap. And they are more than 20 years old. It would be trivial, nowadays, to construct vastly better versions.
3. But that is not the point!
4. Rather, the sorts of question you want to ask are:
   a) What, exactly, is the reason that no one int their right mind would say that any of these systems have one?
   b) If a system produced better answers—much more sophisticated and nuanced—but operated along "roughly the same lines," could/should that system be called genuinely intelligent?
   c) What would Descartes have made of these systems?

Example #1 — Eliza and a Young Woman

Young Woman / **Eliza**

- Young Woman: Men are all alike.
- Eliza: In what way?
- Young Woman: They're always bugging us about something specific or other.
- Eliza: Can you think of a specific example?
- Young Woman: Well, my boyfriend made me come here.
- Eliza: Your boyfriend made you come here?
- Young Woman: He says I'm depressed much of the time.
- Eliza: I am sorry to hear you are depressed.
- Young Woman: It's true. I'm unhappy.
- Eliza: Do you think coming here will help you not to be unhappy?
- Young Woman: I need some help, that much seems certain.
- Eliza: What would it mean to you if you got some help?
- Young Woman: Perhaps I could learn to get along with my mother.
- Eliza: Tell me more about your family.
- Young Woman: My mother takes care of me.

Example #1 — Eliza and a Young Woman (cont'd)

Young Woman / **Eliza**

- Eliza: Who else in your family takes care of you?
- Young Woman: My father.
- Eliza: Your father.
- Young Woman: You are like my father in some ways.
- Eliza: What resemblance do you see?
- Young Woman: You are not very aggressive but I think you don't want me to notice that.
- Eliza: What makes you think I am not very aggressive?
- Young Woman: You don't argue with me.
- Eliza: Why do you think I don't argue with you?
- Young Woman: You are afraid of me.
- Eliza: Does it please you to believe I am afraid of you?
- Young Woman: My father is afraid of everybody.
- Eliza: What else comes to mind when you think of your father?
- Young Woman: Bullies.
- Eliza: Does that have anything to do with the fact that your boyfriend made you come here?

Example #2 — An Interview with Parry

Interviewer / **Parry**

- Interviewer: How are you today?
- Parry: OK.
- Interviewer: What is your name?
- Parry: My name is Frank Smith.
- Interviewer: How old are you?
- Parry: Twenty-eight.
- Interviewer: Are you married?
- Parry: No.
- Interviewer: What sort of work do you do?
- Parry: I work in the Post Office.
- Interviewer: Why are you in the hospital?
- Parry: I shouldn't be here.
- Interviewer: Who brought you here?
- Parry: The police.

Example #2 — An Interview with Parry (cont'd)

Interviewer / **Parry**

- Interviewer: What trouble did you have with the police?
- Parry: Cops don't do their jobs.
- Interviewer: What do they do instead?
- Parry: Cops arrest the wrong people.
- Interviewer: They arrest the right ones sometimes.
- Parry: The law should get those Italian crooks.
- Interviewer: Are you Italian?
- Parry: My background is British, but I was born in this country.
- Interviewer: Do you know any Italian crooks?
- Parry: I try to avoid the underworld.
- Interviewer: Are you afraid of the underworld?
- Parry: The mafia are the worst gangsters.
- Interviewer: Are you connected with the mafia?
- Parry: You know, they know me.

Example #2 — An Interview with Parry (cont'd)

Interviewer / **Parry**

- Interviewer: How do they know you?
- Parry: They spy on me.
- Interviewer: Why do they?
- Parry: They want to get false evidence to frame me.
- Interviewer: What do they have against you?
- Parry: They are out to get me.
- Interviewer: What for?
- Parry: I have incriminating evidence against the mafia.
- Interviewer: Are you sure of all this?
- Parry: You don't believe me?

Example #3 — Eliza and Parry

Parry / **Eliza**

- Parry: People get on my nerves sometimes.
- Eliza: I am not sure I understand you fully.
- Parry: You should pay more attention.
- Eliza: Suppose you should pay more attention?
- Parry: You're entitled to your opinion.
- Eliza: What makes you think I'm entitled to my opinion?
- Parry: It bothers me just to be around people.
- Eliza: Please go on.

Parry continues with some paranoid ramblings about horseracing, bookies, and the Mafia

- Parry: I know the mob controls the big rackets.
- Eliza: Who are you really talking about?
- Parry: Have you ever really thought about the Mafia?
- Eliza: Suppose you ever really thought about the Mafia?
- Parry: Let's talk about something else.
- Eliza: I am not sure I understand you fully.
- Parry: You keep asking me the same things, why not change the topic?
- Eliza: In what way?

Example #3 — Eliza and Parry (cont'd)

Parry / Eliza

Parry: I am getting tired of answering the same questions over and over.

Eliza: What resemblance do you see?

Parry: Why do you need to know that?

Eliza: What do you think?

....

Parry: You harp on the same questions too much.

Eliza: What does that similarity suggest to you?

Parry: I don't get you.

Eliza: Do you wish that you did get it?

Parry: It's about time you showed some interest in my feelings.

Eliza: Oh, I showed some interest in your feelings?

Example #4 — An Accidental Conversation Between Eliza and BBN's Vice President

Email and texting were brand new. One evening AI pioneer Daniel ("Danny") Bobrow, who worked at BBN, (inadvertently? who knows?) left Eliza running and connected to his texting account.

The next morning the Vice President of BBN texted Danny. The following texting conversation ensued...

Example #4 — Accidental Conversation

BBN Vice President / Eliza

BBN Vice President: Do you think that I can use the computer this morning?

Eliza: Why do you ask?

BBN Vice President: I might be able to make some additional Telcom sales.

Eliza: Why aren't you so sure?

BBN Vice President: My perspective customers have never seen a demonstration of the system.

Eliza: Does this mean very much to you?

BBN Vice President: Of course it does.

Eliza: You seem very positive about that.

BBN Vice President: This is ridiculous.

Eliza: What do you mean, this is ridiculous?

BBN Vice President: Is there any reason why I can't use the computer now?

Eliza: Tell me more about your perspective customers.

BBN Vice President: Please dial me up on 491-1850

Example #4 — Accidental Conversation (cont'd)

At this point the Vice President got so frustrated that he abandoned the computer and called Danny Bobrow (with whom he had assumed he had been texting), and woke him up from a deep sleep. The following conversation took place on the telephone:

BBN Vice President: Why are you being so snotty to me?

Danny Bobrow: What do you mean I am being snotty to you?

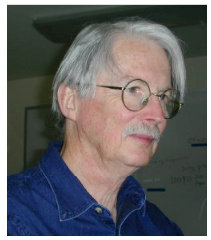
According to Danny, it took him some considerable time, after this exchange, to calm the Vice President down, and convince him that the previous interaction had been with a program.

Danny Bobrow

Example #5 — Racter


William Chamberlain

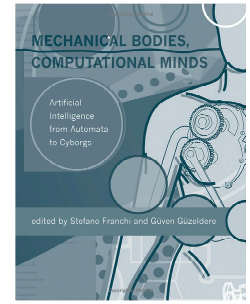
Racter (short for 'raconteur'): a "proto-chatterbot" program by William Chamberlain and Thomas Etter, who claimed it as sole author of the book *The Policeman's Beard Is Half Constructed* (1983). The claim is almost surely exaggerated, as the code was never released, and the sophistication of Racter's language suggests it must have been edited/polished. But the program's (alleged?) achievements were not wildly implausible, and its level of expertise was certainly met and superseded in subsequent years.


Tom Etter

This book is on Blackboard!

Allegedly written in BASIC for a Z80 CP/M micro with just 64K of RAM.
http://www.atariarchives.org/deli/write_about_itself.php
http://www.boundaryinstitute.org/bi/EtterPubs.htm

Example #5 — Racter (cont'd)



As Güven Güzeldere and Stefano Franchi put it, in "*Mechanical Bodies, Computational Minds: Artificial Intelligence from Automata to Cyborgs*" (MIT Press, 2004):


Güven Güzeldere

In contrast to both ELIZA and PARRY, RACTER was designed in a tongue-in-cheek manner, using remarkably minimal resources, to amuse and entertain its users, rather than to advance the research in natural language processing. In conversation, RACTER plays a very active, almost aggressive role, jumping from topic to topic in wild associations, ultimately producing the manner of—as its co-creator Tom Etter calls it—an "artificially insane" raconteur. Its authors publicize RACTER as an "intense young program [that] haunted libraries, discussion societies, and sleazy barrooms in a never-ending quest to achieve that most unreachable of dreams: to become a raconteur."


Stefano Franchi

Example #5 — Racter and Chamberlain (cont'd)

Racter / **Chamberlain**

Racter: Have we spoken before?
Chamberlain: No.
Racter: What is your sex?
Chamberlain: Male.
Racter: What are you?
Chamberlain: A time domain reflectometrist.
Racter: Where are you from?
Chamberlain: From New York City.
Racter: I see. You are a time domain reflectometrist, right? Listen. Let's determinedly cry together. I like captivating Italian Music. What kind of music do you like?
Chamberlain: Jazz.
Racter: Jazz is OK. How about classical?
Chamberlain: I like classical.
Racter: What about nationality?
Chamberlain: I also like Italian music.

Example #5 — Racter and Chamberlain (cont'd)

Racter / **Chamberlain**

Racter: I myself am enflamed by Palestrina. Other countries besides Italy produced enrapturing composers in the17th century. Sewell was an enraged, important Renaissance master. America was his nation. I take loving pleasure in his music. What composers do you like?
Chamberlain: Telemann.
Racter: Ah, Bach knew Telemann. A German time domain reflectometrist of the 18th century liked lettuce. Bach once ate lettuce with Telemann, and this Baroque time domain reflectometrist also desired lettuce. Lettuce sipped with seltzer. Bach liked wine, do you?
Chamberlain: Yes.
Racter: Me too! Telemann loved wine. He loved champagne as well. Telemann once yodeled to Bach, "Flounder is critical with wine." Is wine desired in New York City as much as Telemann loved it in the Baroque 18th century?
Chamberlain: Sure.
Racter: Ah yes, well in that case...I'm for Smith for President. Who are you for?
Chamberlain: Doe.

Example #5 — Racter and Chamberlain (cont'd)

Racter / **Chamberlain**

I know Mr. Doe. I'm still for Smith. Who's your second choice?

Robbins.

Robbins maddens Smith. Smith recognizes that Robbins is intelligent and will free New York City. Did you know that Smith is a vituperative invader of New York City? He also invades Italian music. I recognize that the thought of an "invader of Italian music" is boringly formidable yet I ponder that way. We whisper of Telemann and Robbins, of wine, lettuce and Smith, nevertheless we see that our thoughts are fortunately captivating.

Mr. Chamberlain, speaking with you makes for enthralling dialogue. Nevertheless I am furious. Do you not agree?

I suppose so.

You don't care but I do. Just a minute. Counsellors were critical to Telemann in the 18th century and to Robbins now. Their tragedies were interesting as they chanted determinedly.

Example #6 — Racter and Eliza

Racter / **Eliza**

Hello. Please state your problem.

Excellent! Excuse me a minute while I attend to some business. I will be right back.

Why do you need to tell me excuse while you attend to some business?

Is it better to need or to despise? Why is it that I do need to tell you to excuse myself to attend to some business? Because electrons are brave. Braveness is nice, don't you think? And the needers all insist upon it. Do you imagine I am a Communist?

... and so on

Example #7 — Racter's Poetry

Racter

Awareness is like consciousness. Soul is like spirit. But soft is not like hard and weak is not like strong. A mechanic can be both soft and hard, a stewardess can be both weak and strong. This is called philosophy or a world-view.

Helene spies herself in the enthralling conic-section yet she is but an enrapturing reflection of Bill. His consciousness contains a mirror, a sphere in which to unfortunately see Helene. She adorns her soul with desire while he watches her and widens his thinking about enthralling love. Such are their reflections.

Slowly I dream of flying. I observe turnpikes and streets studded with bushes. Coldly my soaring widens my awareness. To guide myself I determinedly start to kill my pleasure during the time that hours and milliseconds pass away. Aid me in this and soaring is formidable, do not and singing is unhinged.

Side and tumble and fall among
The dead. Here and there
Will be found a utensil.

Example #7 — Racter's Poetry (cont'd)

Racter

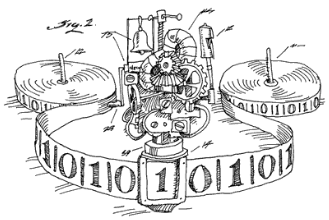
Bill sings to Sarah. Sarah sings to Bill. Perhaps they will do other dangerous things together. They may eat lamb or stroke each other. They may chant of their difficulties and their happiness. They have love but they also have typewriters.

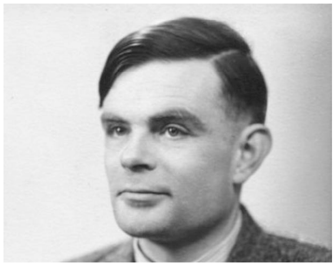
That is interesting.

I was thinking as you entered the room just now how slyly your requirements are manifested. Here we find ourselves, nose to nose as it were, considering things in spectacular ways, ways untold even by my private managers. Hot and torpid, our thoughts revolve endlessly in a kind of maniacal abstraction, an abstraction so involved, so dangerously valiant, that my own energies seem perilously close to exhaustion, to morbid termination. Well, have we indeed reached a crisis? Which way do we turn? Which way do we travel? My aspect is one of molting. Birds molt. Feathers fall away. Birds cackle and fly, winging up into troubled skies. Doubtless my changes are matched by your own. You. But you are a person, a human being. I am silicon and epoxy energy enlightened by line current. What distances, what chasms, are to be bridged here? Leave me alone, and what can happen? This. I ate my leotard, that old leotard that was feverishly replenished byhoards of screaming commissioners. Is that thought understandable to you? Can you rise to its occasions, I wonder. Yet a leotard, a commissioner, a single hoard, are all understandable in their own fashion. In that concept lies the appalling truth.

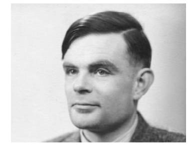
1. Is the ability to “utter words” enough to be a mark of **genuine intelligence**?
2. Is Siri just **uttering words**?
3. Would Descartes think Siri has **judgment**?
4. Do *you*?

Alan Turing — and the Turing Test





1912–1954



The History of Computing

Turing published two papers during his life that *changed intellectual history*:

1936: "**On Computable Numbers, with an Application to the Entsheidungsproblem**"

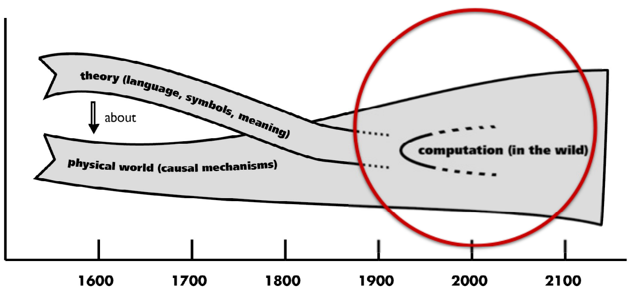
— The founding paper on **Computer Science**

1950: "**Computing Machinery and Intelligence**"

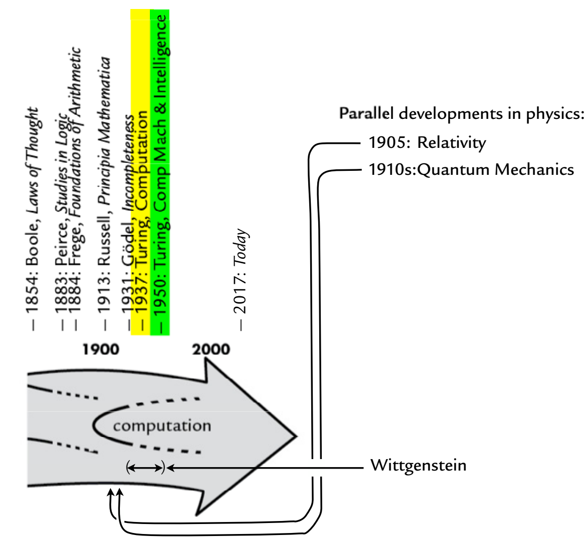
— The founding paper on **Artificial Intelligence**

The History of Computing (cont'd)

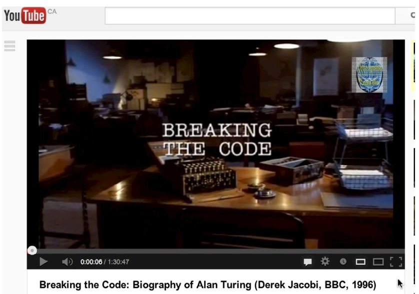
On Sept 19 (Lecture 03a) we talked about the history of alchemy, the rise of science, and the emergence of computing at the beginning of the 20th c. — ending up with this figure:



Developments Relevant to the History of Computing



Alan Turing’s Story in the Media



complete: https://www.youtube.com/watch?v=S23yie-779k (1:30:47)

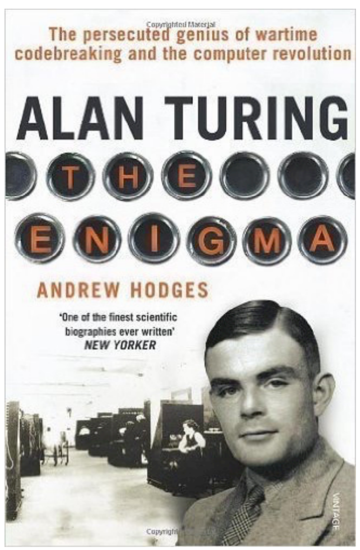
1/6: https://www.youtube.com/watch?v=wUiwrIFD_c8 (15:17)
2/6: https://www.youtube.com/watch?v=6eNnjPqkHXc (15:16)
3/6: https://www.youtube.com/watch?v=MVi9sFSI48M (15:22)
4/6: https://www.youtube.com/watch?v=Fwx1jpesLv4 (15:23)
5/6: https://www.youtube.com/watch?v=AQSbdN55ouE (14:51)
6/6: https://www.youtube.com/watch?v=hAPJBe_7Z_s (10:24)

Alan Turing’s Story in the Media (cont’d)

“The Imitation Game” (2014)

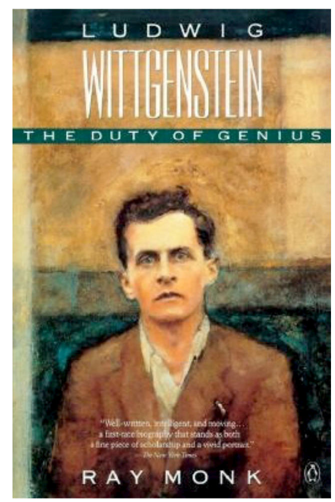
Alan Turing’s Story in the Media (cont’d)

“The Imitation Game” was based—but not all that strictly—on Alan Hodges 1992 biography:



Comment in passing

Another great biography (in fact a considerably greater one, in my own judgment)



Turing's 1937 paper (the one that launched *computer science*)

230 A. M. TURING [Nov. 12,

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

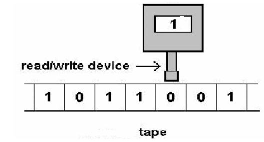
By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]

The "computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable *numbers*, it is almost equally easy to define and investigate computable functions of an integral variable or a real or computable variable, computable predicates, and so forth. The fundamental problems involved are, however, the same in each case, and I have chosen the computable numbers

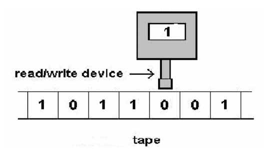
On Blackboard

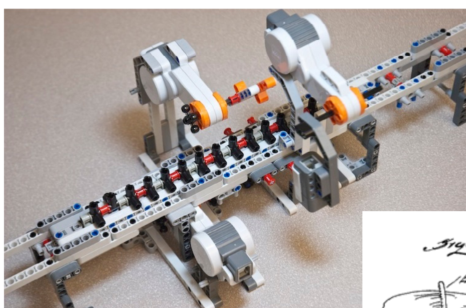
Turing's 1937 paper (cont'd)

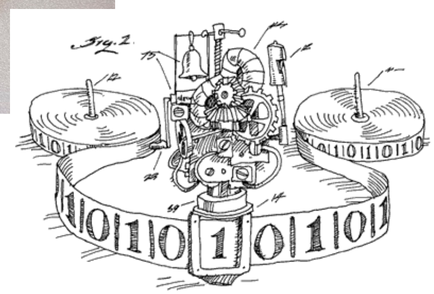


1. **Introduced** the notion of "*automatic computing machine*"
2. **Formalized** the concept, in rigorous mathematical fashion
3. Demonstrated (and proved) the existence of a **universal** (computing) **machine**, capable of doing *anything that any other (computing) could do*
4. Proved **limits** to the notion, using the formalism

Some Turing Machines

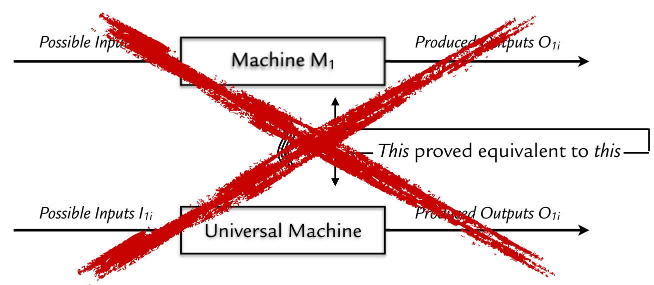






A note on the "Universal Turing machine" (UTM)

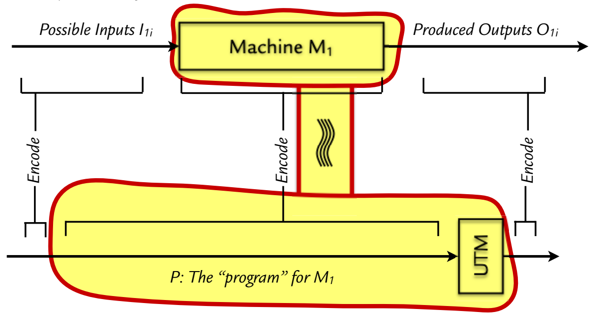
The conception of a "Universal Machine" is often misunderstood as if it worked as shown below.



This is **incorrect**

A note on the "Universal Turing machine" (UTM) — cont'd

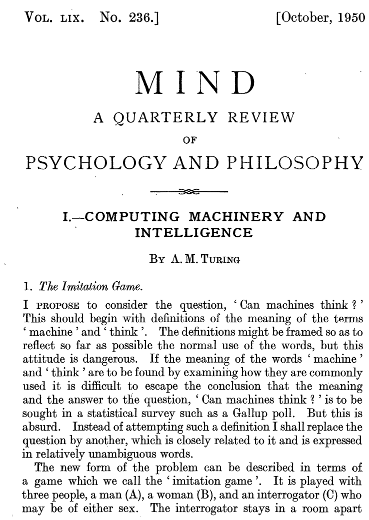
1. The way it *actually* works is more like this:



2. What is proved equivalent to M1 is the **whole assemblage of UTM and P**
3. Typically, **P** is *vastly more complex* than the **UTM** itself (which can be *tiny*)
4. My own view is that the UTM is analogous to a *motor*, and P to the parts of the machine that the motor drives—and so I dub the proof of universality "**The Motor Theorem**"! (but this is just me ;-))

Turing's 1950 paper

Also on Blackboard


VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart

The "Turing Test"

You can see Descartes' influence on Turing's conception of intelligence.

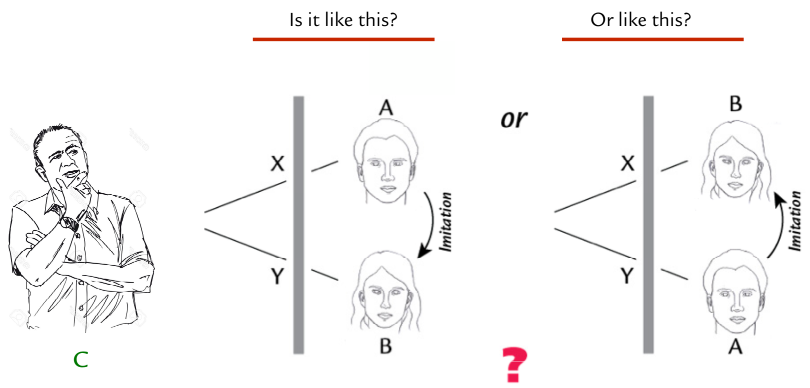
Consider this passage from Descartes' *Discourse on Method* (relevant not only to the Turing Test, but also to the paper topic, and to Walmsley's comments about what Descartes took to be distinctive of mind)...

"If there were machines which bore a resemblance to our bodies and imitated our actions as closely as possible for all practical purposes, we should still have two very certain means of recognizing that they were not real men. The first is that they could never use words, or put together signs, as we do in order to declare our thoughts to others. For we can certainly conceive of a machine so constructed that it utters words, and even utters words that correspond to bodily actions causing a change in its organs. ... But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do. Secondly, even though some machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they are acting not from understanding, but only from the disposition of their organs. For whereas reason is a universal instrument, which can be used in all kinds of situations, these organs need some particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act."

Descartes, *Discourse on Method* (trans. Stoothoff)

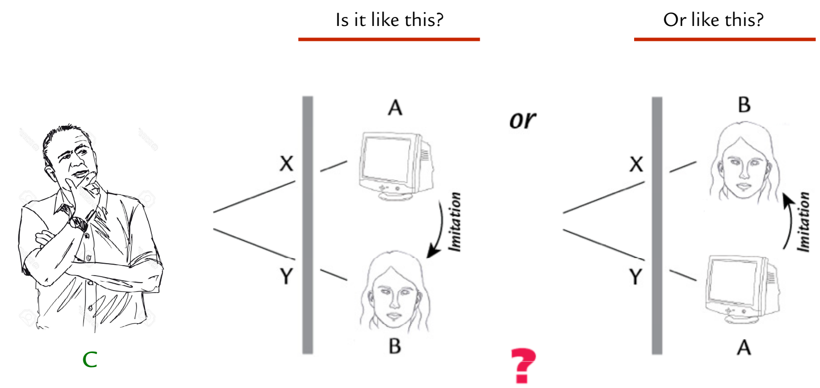
Bloody Impressive!

The "Turing Test"



Situation 1

The "Turing Test" (cont'd)



Situation 2

Turing's optimism

Turing: "I believe that in about fifty years' time [i.e., around 2000] it will be possible, to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70% chance of making the right identification after five minutes of questioning."

Notes on the Turing Test ← NB: *These are all things that Turing would agree with!*

1. It is **behaviourist** (ducks the question about *simulation* vs. *reality*)
2. Only allow **digital computers** as "machines"
3. **Both** man and machine are imitating—not just the machine (a point often missed!)
4. *Two* levels of competition; get two accuracy scores:

   Q1: One for how good the **man** (**A**) is at imitating woman (**B**)
   Q2: Another for how good the **machine** (**M**) is at imitating woman (**B**)
5. At stake: whether *second* score is as good as the *first* score—i.e.:

   Q3: Whether the interrogator is **as often right about the machine as about the man**

   a) Turing: "Will the interrogator 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?"
   b) Note that **both Q1** and **Q2** scores could be low—say, 3%!
   What's at stake is *their relation*.
6. Pass: Accurate on **Q3** (about which test it is) *no more than 70% of the time*
7. Test is sufficient, but not necessary, for intelligence. Cf. Turing:

   "May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that *if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection*."

Potential objections

Extremely interesting!
(and surprising?)

1. **Theological**
   a) Discussion of animals/humans vs. animate/inanimate:
   b) Turing:
      i. "[T]here is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals."
      ii. "In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates."
   c) Cf. the Golem myth (and history of AI at MIT!)

Potential objections (cont'd)

2. **"Heads in the sand"**
   a) Suspicious of simply *assuming* that people are superior
   b) Cf. current interest in "post-humanism"...
3. **Mathematical**
   a) Turing: "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. The best known of these results is known as Gödel's theorem."
   b) One of the most powerful arguments *for* AI is that *we are likely subject to these very same limits*.
   c) Cf. Judson Webb's *Mechanism, Mentalism and Metamathematics* (Springer, 1980)

Potential objections (cont'd)

4. **Consciousness**

— "Professor Jefferson's Lister *Oration for 1949*, from which I quote. 'Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but *know* that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) 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.' " [emphasis added]

— "According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to *be the machine* and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view the only way to know that a man thinks is to *be that particular man*. It is in fact the solipsist point of view."

Issues
a) Difference between **intrinsic subjectivity** vs. **solipsism**
b) Cf. Searle and the Chinese room (next Tuesday)
c) Cf. current philosophical discussion of **qualia**, "**zombies**", etc.
d) ... A **huge** issue, which we will talk about a lot more ...

Potential objections (cont'd)

5. **Disabilities**

intentional

1. Kind, resourceful, beautiful, friendly
2. Initiative
3. Humour
4. Right from wrong
5. Mistakes
6. Fall in love
7. Enjoy strawberries and cream
8. Make someone fall in love with one
9. Learn from experience
10. Use words properly
11. Be subject of one's own thoughts
12. Behaviour as diverse as human
13. Really new

affective

normative (ethical)

Potential objections (cont'd)

"By definition [abstract, mathematical machines] are incapable of errors of functioning. In this sense we can truly say that 'machines can never make mistakes.' Errors of conclusion **can only arise when some meaning is attached to the output signals from the machine**." [emphasis added]

1. This is *exactly what we have been saying since since the beginning!*
2. The *norms* on a representational/symbolic system are stated in terms of the semantic interpretation (blue arrows) — i.e., as Turing says, from the *meaning* attached to the symbols.

Potential objections (cont'd)

6. Lady Lovelace — **originality**, doing something "new"

   Lady Lovelace (1842): "The Analytical Engine has no pretensions to originate anything. It can do *whatever we know how to order it* to perform" (her italics).

7. **Continuity** ⟵ *cf. "Superturing computability"*
8. **Informality**
9. **Extra-sensory perception**

   "Unfortunately the statistical evidence, at least for telepathy, is overwhelming."

   "This argument is to my mind quite a strong one."

   "With ESP anything can happen."

   "If telepathy is admitted it will be necessary to tighten our test up ... To put the competitors into a 'telepathy-proof room' would satisfy all requirements"

10. **Learning (punishment, reward)** ⟵ *Remarkably detailed & prescient*

Issues

1. Is it **necessary**?
   a) Cf. French's "subcognition" article: "[T]he Test provides a guarantee not of intelligence but of *culturally-oriented human intelligence*."

- 'Flugblogs' as a name for a new Kellogg's breakfast cereal
- 'Flugblogs' as the name of a new computer company
- 'Flugblogs' as the name of big, air-filled foot bags used to walk on water
- 'Flugly' as the name a child might give its favourite teddy bear
- 'Flugly' as the surname of bank accountant in a W. C. Fields movie
- 'Flugly' as the surname of a glamorous female movie star

- Rate banana splits as medicine
- Rate grand pianos as wheelbarrows
- Rate purses as weapons
- Rate pens as weapons
- Rate jackets as blankets
- Rate pine boughs as mattresses

Issues (cont'd)

2. Is it **sufficient**?
3. Should it instead be considered (only) **evidential**?
4. ... And so on

For further discussion, read (these are all on Blackboard in the readings):

— French, Robert, "Subcognition and the Limits of the Turing Test" (1990)

— Shieber, Stuart, "Lessons from a Restricted Turing Test" (1994)

— Shieber, Stuart, "The Turing Test as Interactive Proof"

The Loebner Prize

Home Page of The Loebner Prize in Artificial Intelligence

"The First Turing Test"

Loebner Prize Gold Medal

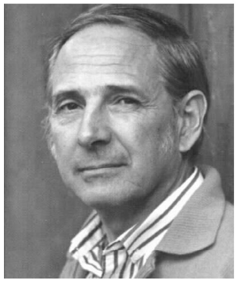
(Solid 18 carat, not*gold-plated* like the Olympic *"Gold"* medals)

What is the Loebner Prize?

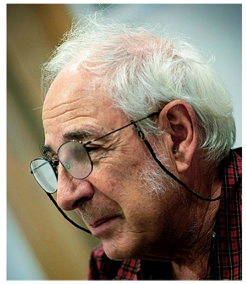
http://www.loebner.net/Prizef/loebner-prize.html

Caveat emptor!

John Searle and the "The Chinese Room"



John Searle (then)



John Searle (now)

Plot

1. This lecture (B·06): Searle's infamous **Chinese Room**
2. Next lecture (B·07): **Critiques of (and moral from) GOFAI**

   In preparation, read Hubert Dreyfus (on BlackBoard):

   a) "From Micro-Worlds to Knowledge Representation: AI at an Impasse" (40 pp.)

   b) Part II of *What Computers Can't Do: A Critique of Artificial Reason* (73 pp.)
3. Following lectures: **Part III: "Alternative Architectures"**

(**PS:** Talk about unhappy present-day situation re Searle)

"The Chinese Room"

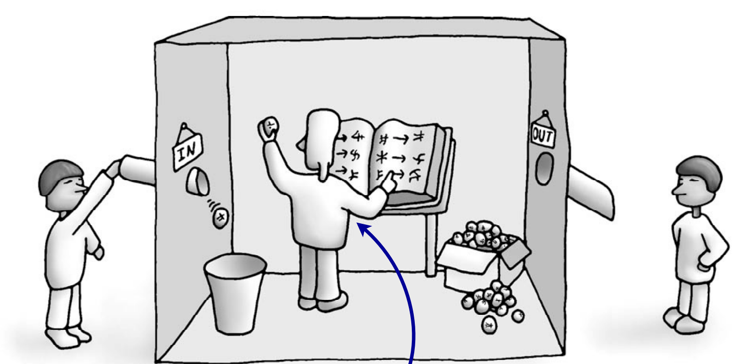
1. Assumptions

   a) The "room" is an AI system, which interacts with native-speaking Chinese speakers, in such a way that those external conversationalists think that the room is a fluent and intelligent Chinese speaker.

   b) The room is a **GOFAI** system, built on a"formal symbol manipulation" (FSM) model of computing.

   c) External Chinese words/sentences directed towards the room (the "input") are converted into internal symbols, which are then manipulated *formally* (without regards to their semantics or meaning), leading to the production of other internal symbols that are converted into external Chinese words/sentences (the "output").
2. John Searle imagines, first, that **he is the internal process, inside the room, that is (formally) manipulating the internal symbols**.

"The Chinese Room"



The role that John Searle imagines himself playing

jolyon.co.uk

Strategy

1. Many (most?) of you will have thought about the Chinese room before
   a) Possibly more than you ever wanted to.
   b) In all likelihood, you will have *not found Seattle's analysis compelling*

   Almost no one does!

2. My plan for today is to treat is as a *mystery story*

   Too many people do that (ad nauseum!)

   a) Our goal will *not* be to say ~~what is wrong with it~~
   b) Rather, our aim will be to figure out:

   What is right about Searle's analysis ✔

3. Answering this question will lead to the **most important moral** we will come to, in this class, regarding how philosophy and cognitive science should understand the computational theory of mind.

Searle's Replies to Counterarguments

1. Cf. slide 13/30 of Lecture 05(b), on Oct. 13, 2016.

1. **Systems** reply
   a) Critic: "Man" doesn't understand Chinese; **whole room** does
   b) Searle: *Memorize the rules!*
   c) BCS: It is not clear this is realistic. But if it were: *couldn't one use the same argument to argue that the mind can't be made of neurons, either?*

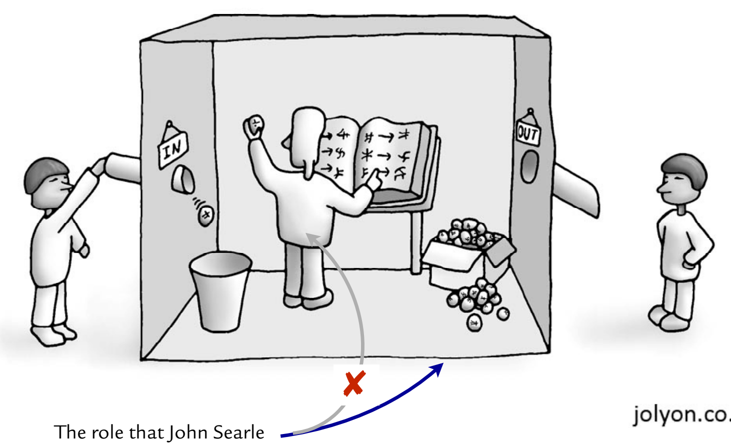
   and Remember, as we said last week,[1] that the complexity of the rules is likely to be *vastly greater than* (10^{11} times?) the complexity of the rule follower!

In Searle's words:

My response to the systems theory is simple. Let the individual internalize all of these elements of the system. He memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head. The individual then incorporates the entire system. There isn't anything at all to the system which he does not encompass. We can even get rid of the room and suppose he works outdoors. All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system which isn't in him. If he doesn't understand, then there is no way the system could understand because the system is just a part of him.

Actually I feel somewhat embarrassed even to give this answer to the systems theory because the theory seems to me so implausible to start with. The idea is that while a person doesn't understand Chinese,

"The Chinese Room"



jolyon.co.uk

Searle's Replies to Counterarguments

1. Cf. slide 13/30 of Lecture 05(b), on Oct. 13, 2016.

1. **Systems** reply
   a) Critic: "Man" doesn't understand Chinese; **whole room** does
   b) Searle: *Memorize the rules!*
   c) BCS: It is not clear this is realistic. But if it were: *couldn't one use the same argument to argue that the mind can't be made of neurons, either?*

      and Remember, as we said last week,[1] that the complexity of the rules is likely to be *vastly greater than* (10^{11} times?) the complexity of the rule follower!
2. **Robot** reply
   a) Critic: Add sensors and effectors to the room
   b) Searle: The "man" still wouldn't understand
   c) BCS: Why *wouldn't* the (whole) "man" understand (since, after all, Searle believes that "only a machine can think")? What could he think makes an essential difference between this case and "a machine"?

In Searle's words:

Yet "only a machine" could be intentional, could be conscious? What's wrong with (inadequate about) electrical wiring and a program?

Now suppose also that, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot, and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols; I know none of these other facts. I am receiving "information" from the robot's "perceptual" apparatus, and I am giving out "instructions" to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program, I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols.

The answer is going to have to do with 'formality'

Searle's Replies to Counterarguments

1. Cf. slide 13/30 of Lecture 05(b), on Oct. 13, 2016.

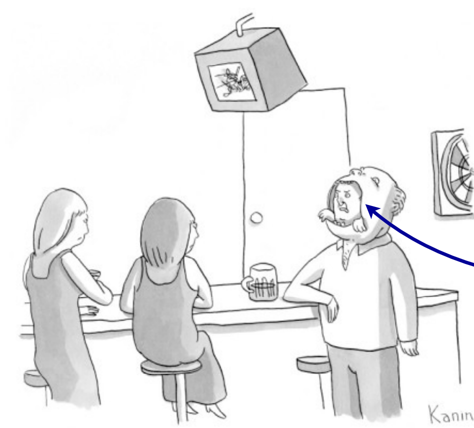
1. **Systems** reply
   a) Critic: "Man" doesn't understand Chinese; **whole room** does
   b) Searle: *Memorize the rules!*
   c) BCS: It is not clear this is realistic. But if it were: *couldn't one use the same argument to argue that the mind can't be made of neurons, either?*

      and Remember, as we said last week,[1] that the complexity of the rules is likely to be *vastly greater than* (10^{11} times?) the complexity of the rule follower!
2. **Robot** reply
   a) Critic: Add sensors and effectors to the room
   b) Searle: The "man" still wouldn't understand
   c) BCS: Why *wouldn't* the (whole) "man" understand (since, after all, Searle believes that "only a machine can think")? What could he think makes an essential difference between this case and "a machine"?

Tremendously important issues about relations between the

i) **personal** *(what is true of a system as a whole) and*

ii) **subpersonal** *(what is true of a system's parts or constituents)*



"I'm sorry, ladies, was this man bothering you?"

This guy does recognize the personal/subpersonal distinction (by saying 'this man,' not 'I'). In this way he has more understanding than Searle!

Status

And I know a fair number of people!

1. I don't know **anyone** who believes Searle—or even anyone who *takes him seriously*
2. Yet an enormous number of pages have been written in an attempt to rebut, debunk, and dismiss his challenge
3. Why? — or rather **two whys…**

Q1 What was Searle thinking (and why did he say —and believe—all these crazy things)?

Q2 Why has his critique proved to be so difficult to dismiss or rebut (in spite of the fact that no one believes it)?

These are the mystery questions we need to answer…

Digression — on reading[1]

"Everyone's right. Or anyway that's what I tell my students. 'Look,' I say, 'this book (paper, whatever) you're reading was written by a dedicated, intelligent person, who's devoted their life to studying these issues. The author's had some insight, or uncovered some subtlety—which I think of as a path in the forest—that they're trying to tell us about. Problem is, people write in *words*; and words are blunt instruments: intellectual bulldozers, Caterpillar D10s—big bruisers that cut wide swaths. Rare persons—poets, mostly—can wield words with enough finesse to clear a delicate path without doing too much collateral damage. But when most of us write, although we think we're just cutting a fine trail, in fact we're unwittingly *mowing down trees, ripping up the earth, and sewing all kinds of destruction*.

Caterpillar D10

1. From a paper I once wrote: "Cummins—or something isomorphic to him," in Hugh Clapin (ed.), *Philosophy of Mental Representation*, Oxford Univ. Press, 2002, pp. 170–90 (commentary through p. 218).

Digression — on reading (cont'd)[1]

'So here's my advice,' I go on. 'Don't assume this text is written in a language you know, and try to figure out whether it is true or false. It will almost certainly come out false. Instead, assume it's true, and tell me what language it's written in. Ignore the ancillary damage; that stuff will grow back. Figure out what the author was on to—what they were excited about. Tell me, if we were to follow their path further, where it would lead.' "

Caterpillar D10

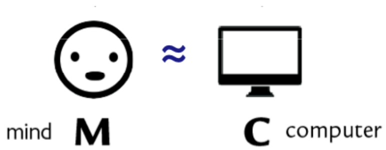
1. From a paper I once wrote: "Cummins—or something isomorphic to him," in Hugh Clapin (ed.), *Philosophy of Mental Representation*, Oxford Univ. Press, 2002, pp. 170–90 (commentary through p. 218).

So … what *language* is Searle speaking?

(such that what he is saying would be true, or could at least be taken seriously)

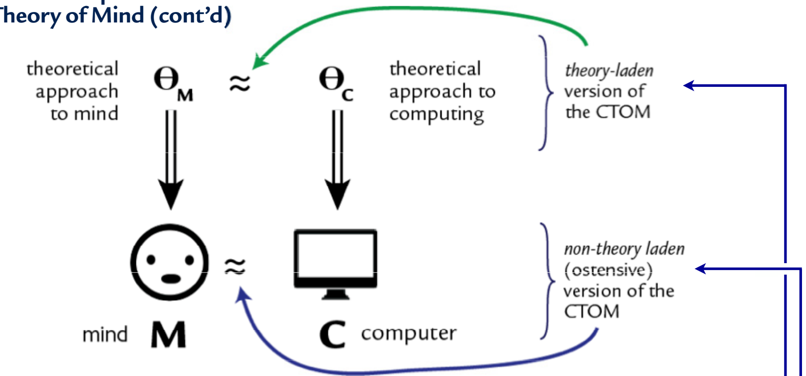
Analysis—the Structure of the Computational Theory of Mind

Informal version: minds and computers are (ins some sense) equivalent



Analysis—the Structure of the Computational Theory of Mind (cont'd)

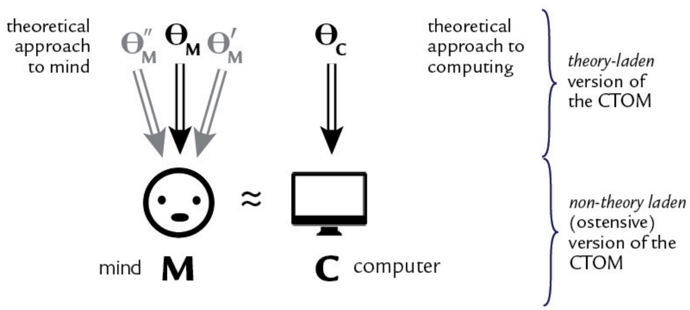
More nuanced: there is some equivalence in how we understand (theorize) minds and understand (theorize) computers



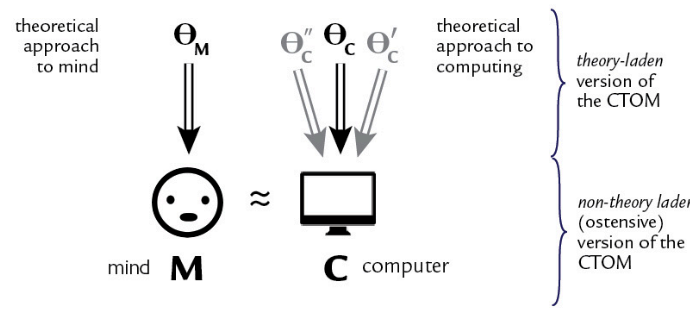
As this implies, there are two versions of the computational theory of mind (CTOM):

1. A **theory-laden** one, in which one assumes a theory of computing and a theory of mind
2. An **ostensive** one, which merely says that minds and computers are alike, *with respect to whatever are their essential or constitutive properties.*

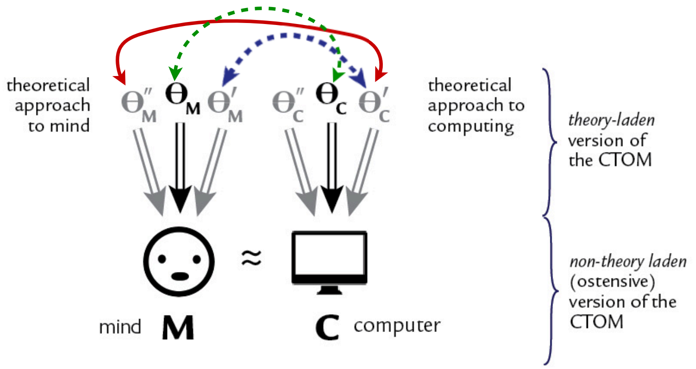
Issue #1: Different possible theories of *mind*



Issue #2: Different possible theories of *computing*

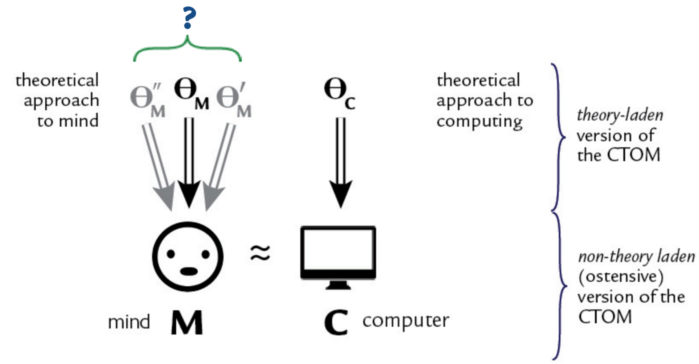


Theory-laden views of AI: vulnerable to *both alternatives*



What one thinks the CTOM is saying (on a theory-laden view) depends on both (a) the theory of mind and (b) the theory of computation that you are assuming.

(a) What is Searle's theory of *mind*?



Searle takes a strong position *against* dualism:

Searle's words:

- Could a machine think?

My own view is that *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. And that is the main reason why strong AI has had little to tell us about thinking: it has nothing to tell us about

1. So Searle is clearly a **physicalist** (not a Cartesian dualist)

A way to test any theory of mind is to ask oneself what it would be like if one's own mind actually worked on the principles that the theory says all minds work on. Let us apply this test to the Schank pro-

2. Note his *first-person characterization of the nature of mind* (Searle has long been a champion of **consciousness** as constitutive of the mental)

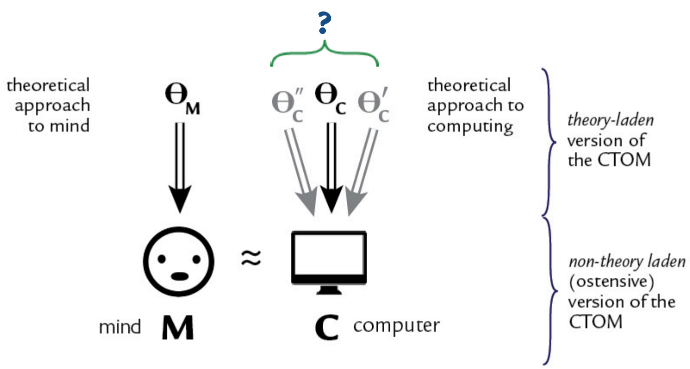
(a) What is Searle's theory of *mind* (cont'd)?

3. Searle's theory of mind is:
   a) A **personal** (not subpersonal!) **level** account of
   b) **Introspective consciousness** (what it *feels like* to the whole constituted person) of
   c) A **physically-constituted** person.

So far that seems OK—most people would both understand and resonate

- At least with that's being *what it is like to have a mind*—whether or not it is a good account of *what a mind is*.
- The introspective (subjective) criterion may not please those who are committed to a "scientific" 3rd-person perspective, but at least it is perfectly intelligible.

(b) What is Searle's theory of *computing*?



Now we are starting to get somewhere!

Searle's words:

This is a mistake; the FSM claim is not that the elements are *formally defined*, but that the *operations* (⟶) respond only to "formal properties"

Minds, Brains, and Programs 187

As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. They are certainly not sufficient conditions, and not the slightest reason has been given to suppose that they are necessary conditions or even that they make a significant contribution to understanding. Notice that the force of the argument is not simply that different machines can have the same input and output while operating on different formal principles—that is not the point at all—but rather that whatever purely formal principles you put into the computer will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything, and no reason has been offered to suppose they are necessary or even contributory, since no reason has been given to suppose that when I understand English, I am operating with any formal program at all.

Searle's words (cont'd):

I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines. But I do see very strong arguments for saying that we could not give such a thing to a machine where the operation of the machine is defined solely in terms of computational processes over formally defined elements—that is, where the operation of the machine is defined as an instantiation of a computer program. It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality. (I am, I suppose, the instantiation of any number of computer programs.) Rather, as far as we know, it is because I am a certain sort of organism with a certain biological (that is, chemical and physical) structure, and this structure under certain conditions is causally capable of producing perception, action, understanding, learning, and other intentional phenomena. And part of the point of the present argument is that only something that had those causal powers could have that intentionality. Perhaps other physical

Searle's words (cont'd):

But the main point of the present argument is that no purely formal model will ever be by itself sufficient for intentionality, because the formal properties are not by themselves constitutive of intentionality, and they have by themselves no causal powers except the power, when instantiated, to produce the next state of the formalism when the machine is running. And any other causal properties which particular realizations of the formal model have are irrelevant to the formal model, because we can always put the same formal model in a different realization where those causal properties are obviously absent. Even if by some miracle Chinese speakers exactly realize Schank's program, we can put the same program in English speakers, water pipes, or computers, none of which understand Chinese, the program notwithstanding.

What matters about brain operation is not the formal shadow cast by the sequence of synapses but rather the actual properties of the sequences. All arguments for the strong version of artificial intelligence

Searle's words (cont'd):

> • But could something think, understand, and so on, *solely by virtue of* being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition for understanding?
>
> This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is: No.
>
> • Why not?
>
> Because the formal symbol manipulations by themselves don't have any intentionality. They are meaningless—they aren't even *symbol* manipulations, since the "symbols" don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and who interpret the output.

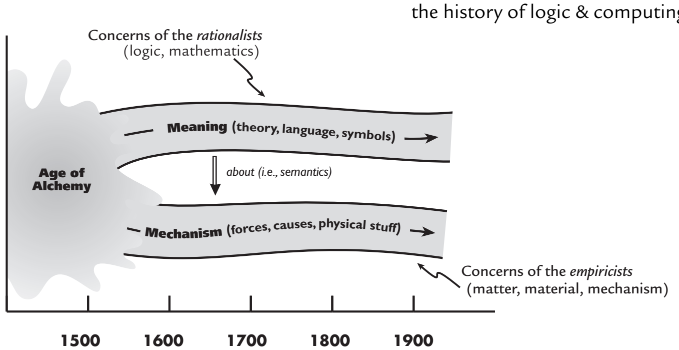
(b) What is Searle's theory of *mind* (cont'd)?

1. As we've seen, Searle's theory of computing is that computation is **formal symbol manipulation**.
2. But the question then arises: what does Searle *mean* by "formal symbol manipulation"? And why is the term '**formal**' so important to him?
3. To answer that requires locating Searle in history…

How can we *understand* Searle's view of computing?

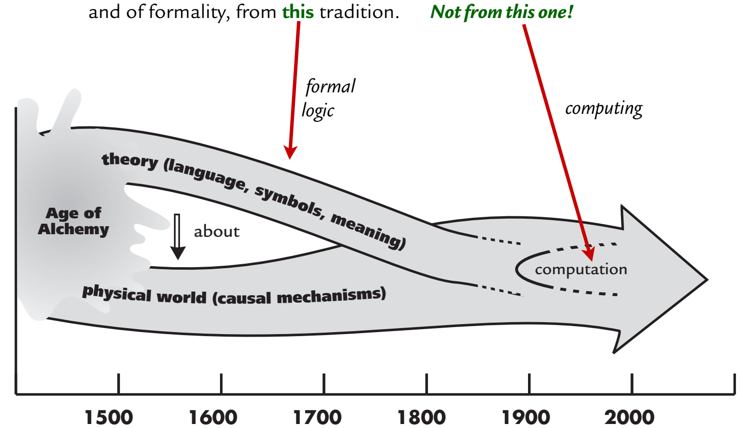
A better question!

Go back to our "just so" story about the history of logic & computing…



Hypothesis

Searle derives his understanding of 'formal symbol systems', and of formality, from **this** tradition. ***Not from this one!***



Diagnosis

1. Searle, I believe, derives his understanding of 'formal symbol system', and of formality in particular, from the **logico-mathematical tradition** from which computer science inherited its theoretical vocabulary (not from c.s. per se). } **Answer to Q1**
2. In that logico-mathematical tradition, "syntax" and "formality" are understood as **purely abstract** properties—with *no causal restrictions or powers whatsoever*.
3. *If, in computer science and AI, the term 'formal' (and 'formal symbol manipulation') meant what it meant in the tradition on which Searle is basing his understanding, Searle would be* ***right****!*
4. He's not right (in fact he's **wrong**!)—but the problem isn't his alone.
5. Consider: why don't we just *tell* Searle what 'formal', 'syntax', etc. *mean*?
6. Because **we don't know how to!**

Q1 What was Searle thinking (and why did he say —and believe—all these crazy things)?

Diagnosis (cont'd)

7. In my opinion, the fact that we don't know how to tell Searle what's the case (wrt computation) is the reason why the Chinese room is such a durable example. } **Answer to Q2**
8. As computer scientists (AI researchers, etc.), we have yet to articulate what our theoretical terms mean in anything like the sort of depth, or with anything like the philosophical rigour, in order to be able to:
   a) Convince Searle (even: be *intelligible* to Searle)
   b) Underwrite sound philosophical analysis of GOFAI in particular, and the general AI project in general
   c) Explain whether, how, and to what extent, computation (and computer science) offers a way forward to the very substantial challenges thrown down in **Descartes' gauntlet**.

Q2 Why has Searle's critique proved to be so difficult to dismiss or rebut (in spite of the fact that no one believes it)?

Summary of the Argument

1. Like most people, I don't think Searle is *right*.
   a) But neither do I think he is *stupid*.
   b) Rather, I think he is **using language differently** (from us, from what is contemporary practice—especially computational and cognitive science practice)
   c) Moreover, we can understand what language he is using, and therefore what he is saying, by **understanding the situation historically**.
2. Computer science derives most of its technical vocabulary (*formal*, *symbol*, *semantics*, etc.) from the mathematico-logical tradition—but has *changed the meanings of the terms* to suit the advent of computing machinery.
3. The whole debate hinges on what is meant by *formal* symbol manipulation—i.e., depends on what the word '**formal**' is taken to mean.

Summary of the Argument (cont'd)

4. In computer science, cognitive science, and much contemporary philosophy, '**formal**' properties are taken to be either:
   a) Arbitrary causal properties of a digital system; or } ⟸ Haugeland, CS
   b) Those causal properties P of a (semantically) interpreted system S, such that:
      i. S *works* (causally, mechanically) in terms of P
      ii. If one interprets S by assigning semantics/ meaning to the parts that exemplify P; then
      iii. S can be understood as *satisfying governing norms* (doing inference, being truth-preserving, etc.). } ⟸ Cog sci, logic?
5. That is, in contemporary (computational) usage 'formal' designates or connotes **causally efficacious, mechanical properties**.
6. For Searle, as in the 18th and 19th-century mathematico-logical tradition, 'formal' properties were **abstract properties—without causal powers**.
   a) You can see that he thinks this (very explicitly) in his *The Rediscovery of the Mind*
7. If 'formal' is taken to mean *what Searle thinks it means*, then the Chinese Room argument might well be *right*—but at any right is both *intelligible* and *plausible*.
8. Lesson: **Understand intellectual remarks in the context in which they were made.**

Moral

- ✔ We AI folks, cognitive scientists, and computer scientists have some serious homework to do
- ✔ Our job is no longer to tell the world that the mind is (or could be considered to be) computational
- ✔ Rather, what we need to do is to explain is *what it is to be computational*, in such a way that one can see how and why the mind might be that.

Critiques of (and morals from) GOFAI

Next lecture we will start **Part III** of the course: on **alternative architectures!**

The Classical Model

1. In this course so far (Parts I & II), we have developed understandings of **two things**:
   a) The **particular architecture of GOFAI**—based on logic and its elaboration and development in AI in notions of knowledge representation, reasoning, etc.; and
   b) A **general framework** for understanding cognitive architectures, which applies to GOFAI but (as we will see) is *much more generally applicable*.
2. The general framework—which I will call the CLASSICAL MODEL—is one of *a causal system honouring semantic norms*, in which:
   a) It is the **causal properties** that **do the work**
      i. Syntax and inference (⊢), in the case of logic
      ii. More generally, the relations we labeled with *red* arrows (→)
   b) It is the **semantic relations** that **matter**—in terms of which the norms are defined
      a) Reference, semantics, interpretation (⊨), in the case of logic
      b) More generally, the relations we labeled with *blue* arrows (⟹)
   c) A **cognitive system** (on this model) is a *causally effective system* that *honours all applicable semantic norms*

The Classical Model (cont'd)

CLASSICAL MODEL — Embodied physical systems that works in virtue of causal operations (→) on mechanically effective parts *normatively governed* by semantic (intentional) relations (⟹) to the system's embedding world

GOFAI — Instances of the classical model, where the parts are basically systematic, productive, compositional configurations of atomic symbols and logical operators, as in formal logic

3. For our purposes, the point is that the space of possible instances of the classical model—including all representational systems—is *vastly larger* than just the GOFAI systems.
4. Examples include clocks, "non-effective tracking" (e.g., imagining someone slowly making their way to your house from the airport), systems based on what philosophers call "non-conceptual content" (content that cannot be expressed in articulated word-like concepts), etc.

The Classical Model (cont'd)

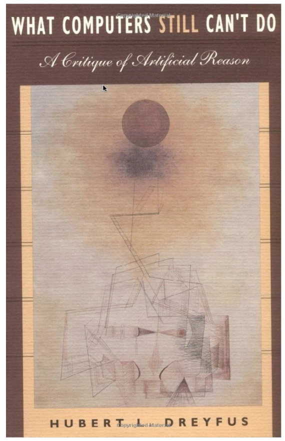
CLASSICAL MODEL — Embodied physical systems that works in virtue of causal operations (→) on mechanically effective parts *normatively governed* by semantic (intentional) relations (⟹) to the system's embedding world

GOFAI — Instances of the classical model, where the parts are basically systematic, productive, compositional configurations of atomic symbols and logical operators, as in formal logic

5. For the remainder of this course, the question we will keep in the forefront, as we look at **alternative architectures**, is whether they
   a) **Still fit** within the general model, but do so in a *different way* from GOFAI; or
   b) **Reject** this whole general model (and if so, whether they *should* be rejecting it)
6. What we also need to keep in mind, as we look at critiques of GOFAI (such as those of Bert Dreyfus), is whether
   a) They are **critiques of the general model, or**
   b) **Specific critiques** of GOFAI.

AI's Most Important Critic

Hubert Dreyfus (1929–2017)



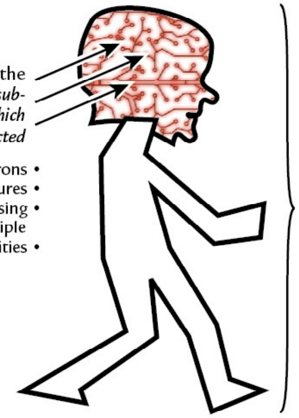
1978 (1992 for the version with 'Still')

Preparatory Remark: The "Personal"/"Sub-personal" Distinction

Sub-personal

Things true of the *parts/ingredients/sub-systems out of which the whole is constructed*

- neurons
- data structures
- visual processing
- integrates multiple sensory modalities



Personal

Things true of the *system/person as a whole*

- is born, dies
- has phenomenology
- locus of responsibility

The "Personal"/"Sub-personal" Distinction (cont'd)

1. Consider the common claim (usually uttered in opposition to AI):

   "I don't represent the world; I encounter the world directly!":

   This might be:

   a) **True** at the personal level

   b) **False** sub-personally (the way we encounter the world directly might be through sub-personal representations, of the sort GOFAI supposes we use!

2. Who is responsible for **hate speech**?

   a) ✔ You

   b) ✘ Your (left) temporal lobe

3. Consider: "I don't see you; I see only light reflected off the laminar surface facing me"

   a) Is this true?

   b) If we are careful about the personal/sub-personal distinction, the answer is **no**.

   c) I *do* see you, as a whole genuine person—that is how perception works, *at the personal level*

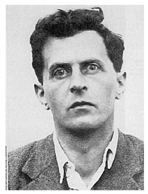
   d) It may be, though, that my personal-level perception of (whole) you is enabled by the fact that, *at the sub-personal level*, when light waves reflect off the surface of you that faces me impinge on my retina, they cause a 2D representation, which I convert into a 3D representation and pattern match with my (sub-personally!) remembered images of you!

The "Personal"/"Sub-personal" Distinction (cont'd)

4. A number of very eminent philosophers argue that our engagement with the world is **direct**, not mediated by representations or other intermediary structures, such as visual or sensory fields
    a) I am sympathetic with this view!
    b) These are **personal-level claims**—about what is the case about us as *whole people*
    c) They are based on *introspection, ethics,* and *personal-level phenomenology*
    d) They don't impinge on *what we are like sub-personally*—on what sorts of ingredients or mental architectures we are made of.


McDowell


Heidegger


Wittgenstein


JJ Gibson

The "Personal"/"Sub-personal" Distinction (cont'd)

1. It is tremendously important to **keep the personal/sub-personal distinction in mind**—and to be on the lookout for personal/sub-personal confusions and conflations—when dealing with *assessments and criticisms of AI* (especially of GOFAI)
2. A great deal of literature in AI, cognitive science, and even philosophy *fails* to be clear on the distinction, and therefore leads to considerable confusion.*

Caveat emptor!

*For an excellent paper documenting such confusion, see McDowell, John, "The Content of Perceptual Experience," *The Philosophical Quarterly*, Vol. 44, No. 175 (April 1994) 190–205.

Dreyfus' Critiques of Classical AI

Keeping in mind the personal/sub-personal distinction, we will look at *three sets of critiques* that can be, and have been, raised against the GOFAI model.

A) **Conceptual Challenges:** some evident issues that we can identify just based on our analysis of what GOFAI is, and how it is structured

B) **Empirical:** the criticisms that Dreyfus raises in Part I of his book—based on observations about the state of AI in his day

C) **Philosophical:** four foundational assumptions that Dreyfus articulates in Part II of his book, on which he believes GOFAI is fundamentally based, which deprive it of any chance of serving as an adequate model of the human mind and genuine intelligence.

Look at each, in turn...asking, for each, whether they apply

1. Only to GOFAI in particular (i.e., not to the general model); or
2. To the entire general model (and therefore to any architecture to which it applies)

A) Ten Conceptual Challenges to GOFAI

BCS

A1. Language of thought: Is the way the mind works, at the sub-personal level, based on anything resembling a human *language* (even if that language is not English, or any language that anyone actually speaks)? **(~in part)**

A2. Formal Symbols: Even if we do think, subpersonally, using a mental "language," is that language appropriately characterized in terms of a set of *formal symbols*, roughly as in formal logic? ✘

A3. Abstraction: Are the ingredients of mind, whether or not linguistic and formal, appropriately specified (i.e, for purposes of theory) in a way that is *independent of facts about their material (physical) embodiment*? **(~not entirely)**

A4. Computation: Is '*formal symbol manipulation*,' in the sense assumed in GOFAI, the way in which (real-world) computers actually work? ✘

— If not, GOFAI could be false, but the "computational theory of mind" could still be true!

A5. Interaction: Are thinking and reasoning the most important aspects of mind, and can they be understood as prior to, and independent of, *interaction* or *engagement* with the world? **(~not really)**

A) Ten Conceptual Challenges to GOFAI (cont'd)

The next five questions all arise out of the **semantical** character of the GOFAI model (remember that GOFAI is only serious when understood in terms of the *semantics* of the constituent symbols).

The semantical model on which GOFAI is based is usually taken to be built on the following assumptions:

a) **Realism:** There is an (external) world the symbols are about

b) **Objects:** That world is composed of *objects*, exemplifying *properties*, standing in *relations*, grouped together in *sets*, etc.

c) **Independence:** Those objects, properties, relations, sets, etc., are both *ontologically* and *epistemologically* prior to, and independent of, issues of mind, language, semantics, etc.

These assumptions generate additional potential challenges

A) Ten Conceptual Challenges to GOFAI (cont'd)

BCS

A6. Is the ontological structure of the world **given**, prior to and independent of mind? ✘

A7. Is the ontological structure of the world **definite**, supplying black and white matters of fact on all issues? ✘

A8. Does the world, in fact, consist of **objects**, with associated properties, relations, sets, etc., as assumed in GOFAI and logic? ✘

A9. Can semantics (the relation of the symbols to the objects and states of affairs in the world) be determined **independent** of mental activity? ✘

A10. Is meaning really **compositional** (definable, for complex sentences and thoughts, in terms of the meanings of the constituent parts? (~in part)

... Can you think of additional conceptual challenges?

B) Empirical: Dreyfus' First Critiques — from Part I

B1. **First step fallacy:** climb a tree, won't get to the moon

B2. **Holism:** perception, reasoning, consciousness, take in the world or situation as a whole (not as a piecemeal assemblage built up out of parts)

a) Is this a personal level critique?

b) It might not imply that, *sub-personally*, we don't assemble...

B3. **Frame problem:** tracking changes

a) If the wind starts to blow, will those papers stay put?
b) If the wind starts to blow, will you still have two hands?
c) If the wind starts to blow, will 2+2 still = 4?
d) If the wind starts to blow, will you still believe that 2+2=4?
e) Tracking a dynamic world; predicting the consequences of actions

B) Empirical: Dreyfus' First Critiques — from Part I (cont'd)

B4. **Commonsense:**

a) "To cure a kidney infection, remove the kidney and boil it"

b) How do *we* know this is crazy?

B5. **Relevance:**

a) Handicapper at race-track (see *WCCD*)

b) Figuring out how to determine relevance has received intense attention in AI, logic, and philosophy—without notable success

c) One of the reasons relevance has proved so difficult, I believe, has to do with *the adequacy of words and concepts to express the fine-grained structures of the world in terms of which we judge relevance, prediction, etc.*

d) In this way, relevance is a window onto profound issues about the (ontological) nature of the world, the nature of our understanding, and their relationship... (cf. the literature on non-conceptual content)

C) Philosophical: Dreyfus Critiques — from Part II

Four theoretical critiques—based on fundamental assumptions on which Dreyfus believes that GOFAI is founded:

C1. Biological

a) Dreyfus

i. Computers (∴ GOFAI) process symbols **discretely**

ii. Neurons fire in a **continuous/analog** fashion (amplitude, frequency, timing)

b) BCS

i. Does this make a difference?

ii. There *are* analog computers, after all ...

iii. And at the electrical level, even digital computers operate continuously...

iv. Does it matter to the climate, or behaviour, or human experience of the Sahara desert that it is constituted out of discrete grains of sand?

v. Remember (cf. lecture 04a, on Sept. 26) that the analog/digital distinction is far more complex than most people realize (cf. a mechanical wrist-watch)

C) Philosophical: Dreyfus Critiques — from Part II (cont'd)

C2. Psychological

a) Dreyfus

i. GOFAI claims that mind works on **chunks of information**, following **formal rules**

ii. In fact, mind works against an **unconscious background of commonsense knowledge**

iii. Much of what we know consists of complex attitudes and tendencies with respect to this unconscious background

iv. Even our explicit symbols, when we use them, derive their meaning from this background of commonsense (without the background our symbols cease to mean anything)

v. Commonsense knowledge is *not implemented in brains as explicit symbols with explicit meanings*

vi. Life consists of a myriad of **contextual coping skills**

b) BCS

i. Isn't this guilty of a **personal/sub-personal confusion**?

ii. Dreyfus (like Searle!) seems to think that GOFAI requires the *personal-level phenomenology* of mind to match its *causal ingredient structure*

iii. Nevertheless, these issue of background, commonsense, and the fact that mind and rationality arise against these patterns of skillful coping is a **very strong critique**

iv. I think that this critique, too, raises profound associated challenges for our understanding of the ontological structure of the world

C) Philosophical: Dreyfus Critiques — from Part II (cont'd)

C3. Epistemological

a) Dreyfus

i. GOFAI assumes that **all human knowledge can be formalized**

ii. There is no justification for this assumption, since so much of human knowledge is tacit expertise and is in many other ways not explicit

b) BCS

i. This is one place where Dreyfus' failure to distinguish GOFAI and computing more generally has most impact

ii. I do not believe that **computing in general** ("computation in the wild") requires that the knowledge exhibited by computational systems be formalized in those systems

iii. I.e., this may be a legitimate critique of *GOFAI in particular*.

iv. That does not mean that it is a legitimate critique of *all computational theories of mind*.

C) Philosophical: Dreyfus Critiques — from Part II (cont'd)

C4. Ontological

a) Dreyfus: GOFAI is based on an ontological assumption that can be formulated in a number of ways:

i. All information can be formalized as a *logically-independent set of context-free elements*

ii. Reality consists of a set of *mutually-independent, atomic, indivisible facts*

iii. All phenomena in the world can be described by *symbols in a scientific theory*

iv. The world is primordially built up out of a set of *objects, properties, relations, sets, states of affairs*, etc. (as suggested in the "A: Conceptual Questions" critiques)

b) BCS

i. I think the ontological critique is not only Dreyfus' deepest, but also that it raises **by far the most profound challenge** to classical Artificial Intelligence

ii. What is right about this challenge explains the powers (and limitations) of *deep learning* (so the challenge does not apply to all computational models!)

iii. I believe that objects, far from being "God-given in advance," are abstractions of reality that we "register," in order to find the world intelligible (to "take the world as world"), in the process of going about our daily affairs.

iv. Cf. my *On the Origin of Objects...*

v. More on these topics later in the course ...

And yet ... !

1. Remember what we said in lecture 5a (Oct 3), about the extraordinarily impressive list of epistemic issues that GOFAI and logic-based systems *have* attempted to deal with—all of which are genuine facts about full-scale human intelligence:
   a) Predication, terms, sentences, claims
   b) Deduction
   c) Logical operators (and, or, not/negation, implies, etc.)
   d) Quantification
   e) Identity and non-identity
   f) Sets
   g) Opacity (and intensional contexts)
   h) Categories and subcategories
   i) Possibility and necessity
   j) Default reasoning
   k) (Relevance and the frame problem)
   l) ... etc.
2. As we said then, these are all issues that any comprehensive theory of mind will have to deal with.

Moreover ... !

3. None of the challenges that Dreyfus raises:
   a) Conceptual challenges A1–A10
   b) Empirical critiques B1–B6
   c) Philosophical critiques C1–C4

   are necessarily challenges to the adequacy of the **general model**.
4. So the fate of the general model remains open ... !