## Part III — Alternative Architectures

# Mental Architecture





## Consider two architectural proposals for mind

1. **GOFAI**
   - a) *Knowledge:* Set of language-like symbolic expressions, formally (causally) manipulated but governed by semantic norms.
   - b) *Thinking:* (Interior, sub-personal) process defined over these expressions (separation of active and passive components)
   - c) *Scale:* 100 *million* facts? (CYC)
2. **Neuroscience**
   - a) *Knowledge:* Arrangements of interconnected neurons
   - b) *Thinking:* Patterns of activation propagated by neurons (no distinction between passive & active components)
   - c) *Scale:* 100 *billion* neurons; 100 *trillion* interconnections!

**Some immediate questions:**

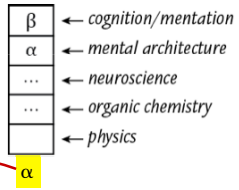
1. Why *these two* architectures?
2. What other architectures are possible?
3. What is an **architecture**, anyway?





## How Does an "Architecture" Relate to a "Mark of the Mental"?

1. A proposed "architecture of mind" α is plausible, wrt to a designated set of mark(s) of the mental β, just in case:
   - **C1** Instances $\alpha_i$ ($\alpha_1$, $\alpha_2$, etc.) of architecture α are capable of exhibiting mark(s) of the mental β, and
   - **C2** A system's exemplification of (marks of the mental) β is *explained in virtue of* its being an instance of α.




2. For example, someone (such as Fodor?) might argue that:
   - a) Logic-based formal symbol manipulation (FSM) is a plausible architecture for mind…
   - b) Wrt to "reasoning and language use" being marks of the mental …
   - c) Because systematicity, productivity, and compositionality are crucial properties of rationality and language use, and …
   - d) One can see how, *in virtue of being a logic-based formal symbol manipulating machine*, a system can exemplify those properties of systematicity, productivity, and compositionality.
3. I.e., α explains β
4. Notes
   - a) It is because of criterion **C2** that the neural level is *not necessarily a plausible architecture of mind*, even though it is obvious that criterion **C1** is satisfied
   - b) According to this definition, architectures for the mind are inherently **sub-personal** (they have to do with how we are made, not with us as whole persons).





## What is an Architecture?

1. At a minimum, an architecture is:
   - a) A physical/mechanical **configuration** or **organization** of system ingredients
   - b) Understood at a certain level of **abstraction/idealization**
   - c) Generally supporting a large set of **different instances**
   - d) (Capable of) exhibiting a certain set of **properties**

   } *To be an architecture is to be a type of mechanical system*
2. More specifically, to specify an **architecture** is to identify
   - a) The space of *possible ingredient or component types*
   - b) The space of *possible ways in which these ingredients can be fitted together and organized*
   - c) The set of *effective transitions*, whereby one configuration of any system (that is an instance of this architectural type) can effectively transition into another one of the other possible configurations
3. Or to put it another way: an architecture is (or encapsulates) the **conceptual design** and **fundamental operational structure** of an *effective mechanical system*
4. For a computational system, or a system of what we are calling the "general model," to be a complete a specification of an architecture should also specify the **semantics**

## Types of architectures of mind—suggestions for how mind is "implemented"

Let's use a triple arrow ('⟹') for "reduces to" or "is architecturally implemented by")

Mind

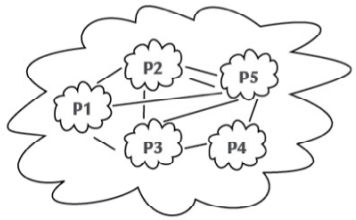
Start with some common forms of **computational implementation**





## "Parallel" Architectures (Multiple Internal Processes)

Mind

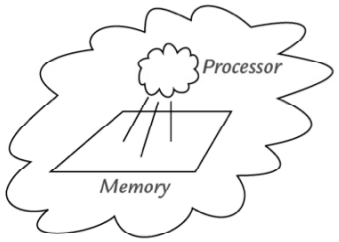







## "Serial" Architectures (Single Internal Process Plus Memory)

Mind



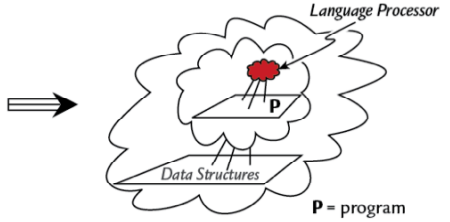




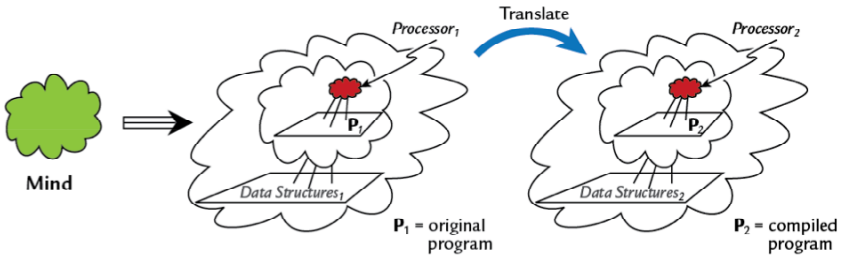

## Classic Programming: A *Double Serial Implementation*

*(also called the "interpreter", or if the program is written in machine language, the "CPU")*

Mind



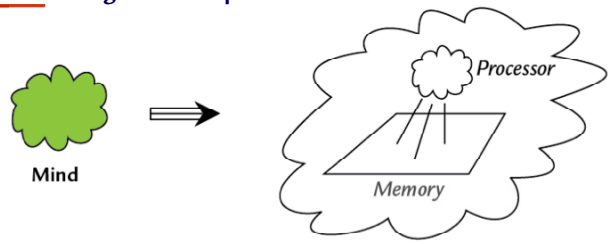
## Compilation: Translation of a program ($P_1$) in one language into an equivalent program ($P_2$) in another








## GOFAI: A *Single* Serial Implementation

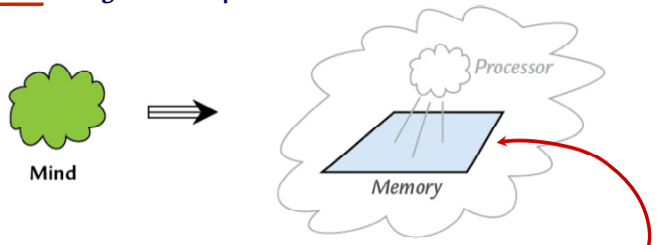



NB, for the computationally sophisticated: The reason for pointing out, in the previous two slides, that standard programming involves a *double implementation* is to contrast that normal "programming" scenario with GOFAI, which is based on only a single one. That is: GOFAI's thesis is that a language-like "mentalese" constitutes something like the *data structures* of the mind—the passive memory-like structures that encode our knowledge of the world. Strikingly, GOFAI is silent on anything that a programmer would call a *program*—any specification of *the procedures that make use of (and modify) those representations.* That is, to put it a bit informally, GOFAI talks about *what we think with*, but doesn't say anything about *how we think*.





## GOFAI: A *Single* Serial Implementation




What does the **memory** (i.e., the data structures) look like in a GOFAI system?

NB, for the computationally sophisticated: The reason for pointing out, in the previous two slides, that standard programming involves a *double implementation* is to contrast that normal "programming" scenario with GOFAI, which is based on only a single one. That is: GOFAI's thesis is that a language-like "mentalese" constitutes something like the *data structures* of the mind—the passive memory-like structures that encode our knowledge of the world. Strikingly, GOFAI is silent on anything that a programmer would call a *program*—any specification of *the procedures that make use of (and modify) those representations.* That is, to put it a bit informally, GOFAI talks about *what we think with*, but doesn't say anything about *how we think*.





## Suggestion #1: The memory would look like a list of expressions in logic

(what people usually imagine)

$$\forall x\,[\,\mathrm{Person}(x) \supset [\,\exists y\,\mathrm{Father}(y,x) \wedge \exists z\,\mathrm{Mother}(z,x)\,]\,]$$

$$\forall x,y\,[\,\mathrm{Uncle}(x,y) \supset [[\,\exists z\,[\,\mathrm{Father}(z,y) \wedge \mathrm{Brother}(x,z)\,]\,] \vee [\,\exists z\,[\,\mathrm{Mother}(z,y) \wedge \mathrm{Brother}(x,z)\,]\,]]]$$

$$\forall x,y\,[\,\mathrm{Aunt}(x,y) \supset [[\,\exists z\,[\,\mathrm{Father}(z,y) \wedge \mathrm{Sister}(x,z)\,]\,] \vee [\,\exists z\,[\,\mathrm{Mother}(z,y) \wedge \mathrm{Sister}(x,z)\,]\,]]]$$

| | |
|---|---|
| $\forall x\,[\,\mathrm{Mother}(x) \supset \mathrm{Female}(x)\,]$ | Mother(Ariadne,Teri) |
| $\forall w\,[\,\mathrm{Father}(w) \supset \mathrm{Male}(w)\,]$ | Father(Llewelyn,Teri) |
| $\forall x,y\,[\,\mathrm{Brother}(x,y) \supset \mathrm{Male}(x)\,]$ | Uncle (Dylan,Teri) |
| $\forall x,y\,[\,\mathrm{Sister}(x,y) \supset \mathrm{Female}(x)\,]$ | Brother(Dylan, Llewelyn) |

## Suggestion #2: No Need for those English Words

$$\forall x\,[\,\text{Person}(x) \supset [\,\exists y\,\text{Father}(y,x) \wedge \exists z\,\text{Mother}(z,x)\,]\,]$$

$$\forall x,y\,[\,\text{Uncle}(x,y) \supset [[\,\exists z\,[\,\text{Father}(z,y) \wedge \text{Brother}(x,z)\,]\,] \vee [\,\exists z\,[\,\text{Mother}(z,y) \wedge \text{Brother}(x,z)\,]\,]]]$$

$$\forall x,y\,[\,\text{Aunt}(x,y) \supset [[\,\exists z\,[\,\text{Father}(z,y) \wedge \text{Sister}(x,z)\,]\,] \vee [\,\exists z\,[\,\text{Mother}(z,y) \wedge \text{Sister}(x,z)\,]\,]]]$$

| | |
|---|---|
| $\forall x\,[\,\text{Mother}(x) \supset \text{Female}(x)\,]$ | Mother(Ariadne,Teri) |
| $\forall w\,[\,\text{Father}(w) \supset \text{Male}(w)\,]$ | Father(Llewelyn,Teri) |
| $\forall x,y\,[\,\text{Brother}(x,y) \supset \text{Male}(x)\,]$ | Uncle (Dylan,Teri) |
| $\forall x,y\,[\,\text{Sister}(x,y) \supset \text{Female}(x)\,]$ | Brother(Dylan, Llewelyn) |

**... etc.**





## Suggestion #2: No Need for those English Words

$$\forall x\,[\,G0021(x) \supset [\,\exists y\,G0349(y,x) \wedge \exists z\,G1172(z,x)\,]\,]$$

$$\forall x,y\,[\,G4421(x,y) \supset [[\,\exists z\,[\,G0349(z,y) \wedge G0629(x,z)\,]\,] \vee [\,\exists z\,[\,G1172(z,y) \wedge G0629(x,z)\,]\,]]]$$

$$\forall x,y\,[\,G0551(x,y) \supset [[\,\exists z\,[\,G0349(z,y) \wedge G0724(x,z)\,]\,] \vee [\,\exists z\,[\,G1172(z,y) \wedge G0724(x,z)\,]\,]]]$$

| | |
|---|---|
| $\forall x\,[\,G1172(x) \supset G0922(x)\,]$ | G1172(Ariadne,Teri) |
| $\forall w\,[\,G0349(w) \supset G0883(w)\,]$ | G0349(Llewelyn,Teri) |
| $\forall x,y\,[\,G0629(x,y) \supset G0883(x)\,]$ | G4421 (Dylan,Teri) |
| $\forall x,y\,[\,G0724(x,y) \supset G0922(x)\,]$ | G0629(Dylan, Llewelyn) |





## Suggestion #3: No Need for the English *Names*, either

$$\forall x\,[\,G0021(x) \supset [\,\exists y\,G0349(y,x) \wedge \exists z\,G1172(z,x)\,]\,]$$

$$\forall x,y\,[\,G4421(x,y) \supset [[\,\exists z\,[\,G0349(z,y) \wedge G0629(x,z)\,]\,] \vee [\,\exists z\,[\,G1172(z,y) \wedge G0629(x,z)\,]\,]]]$$

$$\forall x,y\,[\,G0551(x,y) \supset [[\,\exists z\,[\,G0349(z,y) \wedge G0724(x,z)\,]\,] \vee [\,\exists z\,[\,G1172(z,y) \wedge G0724(x,z)\,]\,]]]$$

| | |
|---|---|
| $\forall x\,[\,G1172(x) \supset G0922(x)\,]$ | G1172(Ariadne,Teri) |
| $\forall w\,[\,G0349(w) \supset G0883(w)\,]$ | G0349(Llewelyn,Teri) |
| $\forall x,y\,[\,G0629(x,y) \supset G0883(x)\,]$ | G4421 (Dylan,Teri) |
| $\forall x,y\,[\,G0724(x,y) \supset G0922(x)\,]$ | G0629(Dylan,Llewelyn) |

**... etc.**





## Suggestion #3: No Need for the English *Names*, either

$$\forall x\,[\,G0021(x) \supset [\,\exists y\,G0349(y,x) \wedge \exists z\,G1172(z,x)\,]\,]$$

$$\forall x,y\,[\,G4421(x,y) \supset [[\,\exists z\,[\,G0349(z,y) \wedge G0629(x,z)\,]\,] \vee [\,\exists z\,[\,G1172(z,y) \wedge G0629(x,z)\,]\,]]]$$

$$\forall x,y\,[\,G0551(x,y) \supset [[\,\exists z\,[\,G0349(z,y) \wedge G0724(x,z)\,]\,] \vee [\,\exists z\,[\,G1172(z,y) \wedge G0724(x,z)\,]\,]]]$$

| | |
|---|---|
| $\forall x\,[\,G1172(x) \supset G0922(x)\,]$ | G1172(F0614, F0258) |
| $\forall w\,[\,G0349(w) \supset G0883(w)\,]$ | G0349(F0774, F0258) |
| $\forall x,y\,[\,G0629(x,y) \supset G0883(x)\,]$ | G4421 (F0532, F0258) |
| $\forall x,y\,[\,G0724(x,y) \supset G0922(x)\,]$ | G0629(F0532, F0774) |

## Suggestion #4: And No Need for the Variables to be *Lexical*

$$\forall x\,[\ G0021(x) \supset [\ \exists y\ G0349(y,x) \wedge \exists z\ G1172(z,x)\ ]\ ]$$

$$\forall x\,\forall y\,[\ G4421(x,y) \supset [[\ \exists z\,[\ G0349(z,y) \wedge G0629(x,z)\ ]\ ] \vee [\ \exists z\,[\ G1172(z,y) \wedge G0629(x,z)\ ]\ ]\ ]\ ]\ ]$$

$$\forall x\,\forall y\,[\ G0551(x,y) \supset [[\ \exists z\,[\ G0349(z,y) \wedge G0724(x,z)\ ]\ ] \vee [\ \exists z\,[\ G1172(z,y) \wedge G0724(x,z)\ ]\ ]\ ]\ ]\ ]$$

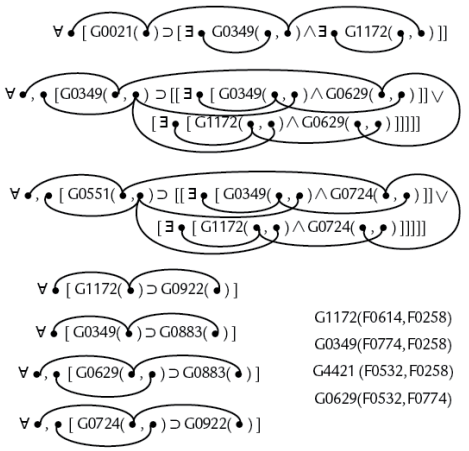
| | |
|---|---|
| $\forall x\,[\ G1172(x) \supset G0922(x)\ ]$ | G1172(F0614, F0258) |
| $\forall w\,[\ G0349(w) \supset G0883(w)\ ]$ | G0349(F0774, F0258) |
| $\forall x\,\forall y\,[\ G0629(x,y) \supset G0883(x)\ ]$ | G4421 (F0532, F0258) |
| $\forall x\,\forall y\,[\ G0724(x,y) \supset G0922(x)\ ]$ | G0629(F0532, F0774) |

**... etc.**





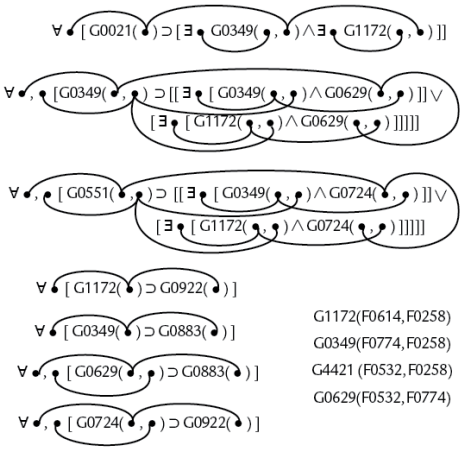
## Suggestion #4: And No Need for the Variables to be *Lexical*








## Suggestion #4: And No Need for the Variables to be *Lexical*




Multiple occurrences of the same name (multiple *instances* of the same name *type*) is merely a **lexical device** to indicate "sameness".

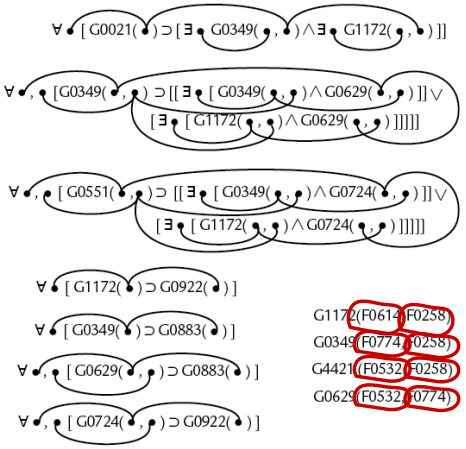
Internally (in the memory) one can indicate this *more directly* (even with actual *coöccurrence*— indicated here with *wires*, but at the level of the architecture it could be *coincidence* [next slide])






**An illustration of *token coöccurrence***

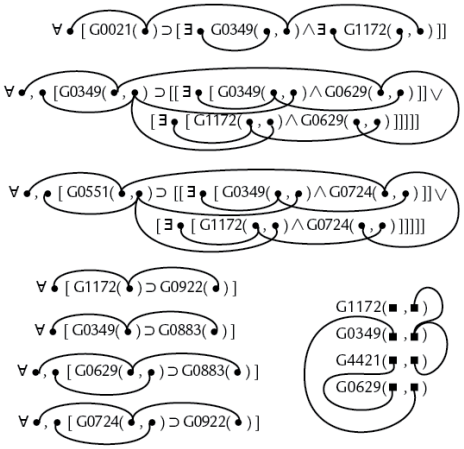
## Suggestion #5: Similarly for *Proper Names*







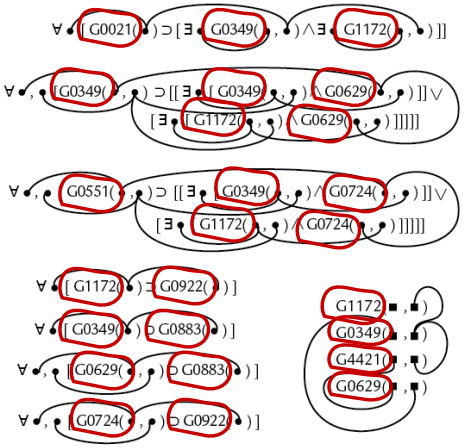

## Suggestion #5: Similarly for *Proper Names*







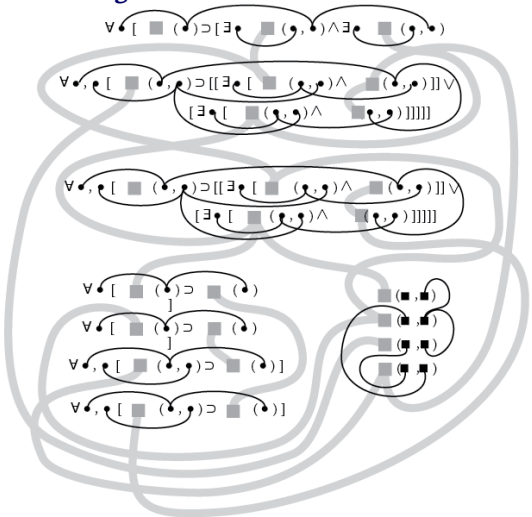

## Suggestion #6: Same thing for the *Predicates*



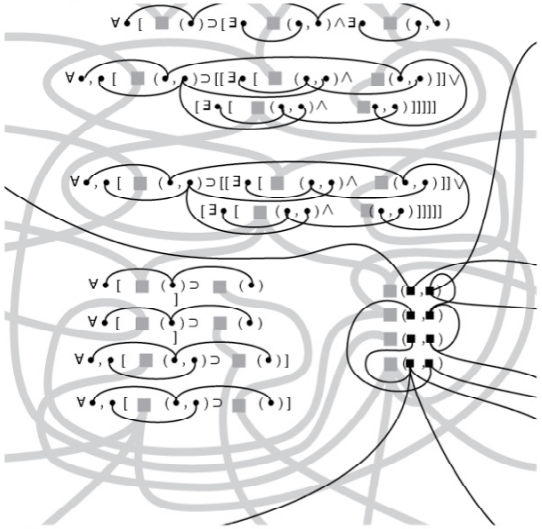





## Suggestion #6: Same thing for the *Predicates*



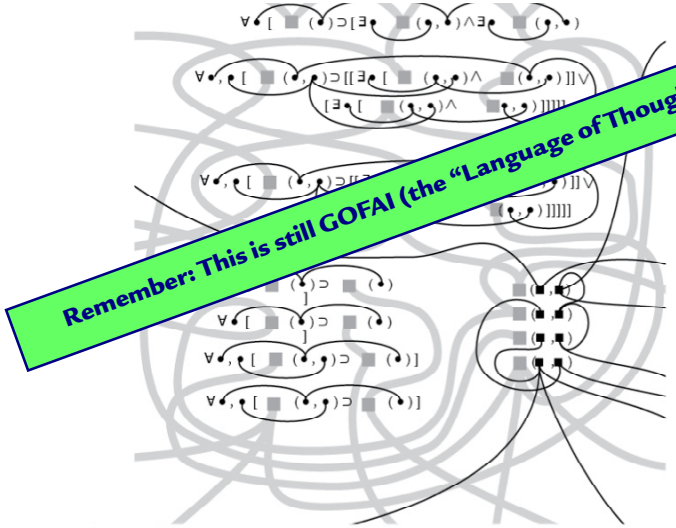
## Suggestion #7: Recognize that it all part of a larger network …







## Suggestion #7: Recognize that it all part of a larger network …



**Remember: This is still GOFAI (the "Language of Thought")!**





## Moral

1. What differentiates one architecture from another is (in general) **not a local fact**, about what a small or atomic item looks like.
2. Rather, architecture has to do with the **global, relational properties** of the parts —how they are grouped, strategies for dealing with them, etc.
3. *This is why simply looking at neurons doesn't reveal the architecture of the human mind*.
4. Looking at interconnections (synapses, axons, neuronal interconnection diagrams, etc.) is a help; but even that is not enough.
5. We would need a **complete, dynamic account of the structured workings** in order to understand how the local structures undergird it.





## Five types of "high-level" mental architecture

1. Language-like
   - **a)** **Language of thought** (GOFAI)
   - **b)** **Poetry**
2. Picture like
   - **a)** **Images**
   - **b)** **Maps**
3. **Procedures**

## Language of Thought

( … we've just seen this one … )





## Poetry

*The fascination of what's difficult*
*Has dried the sap out of my veins, and rent*
*Spontaneous joy and natural content*
*Out of my heart. There's something ails our colt*
*That must, as if it had not holy blood*
*Nor on Olympus leaped from cloud to cloud,*
*Shiver under the lash, strain, sweat and jolt*
*As though it dragged road-metal. My curse on plays*
*That have to be set up in fifty ways,*
*On the day's war with every knave and dolt,*
*Theatre business, management of men.*
*I swear before the dawn comes round again*
*I'll find the stable and pull out the bolt.*

**William Butler Yeats**

The "colt" in this poem ("The Fascination of What's Difficult") is undoubtedly Pegasus, the mythical winged horse of poetic inspiration. This poem was written when Yeats was enmeshed in (and frustrated by!) the foundation and managing of the Abbey Theatre, and in fighting its political and financial battles.



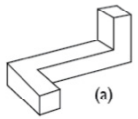


## Images





## Images (cont'd)



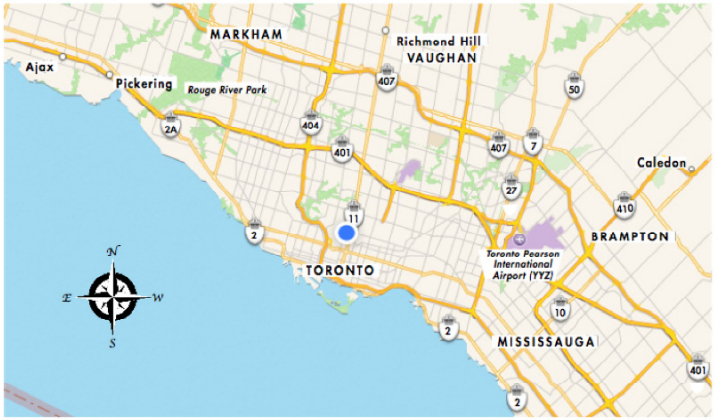

Which of (b), (c), and (d) are possible orientations of the figure given in (a)?

## Maps






## Maps cont'd — A version for the ceiling of the Dentist's office!










## Architectural" Dimensions for Evalution

*The issues one needs to find out about, for any proposed architecture...*

1. **Statics**
   a) What are they *made of, structurally*?
   b) How are parts *assembled*, their forms of *composition*?
2. **Dynamics**
   a) How are they *used*?
   b) What *processes* are defined over or in terms of them?
3. **Semantics (interpretation) and norms**
   a) How do they *represent* the world?
   b) What is it for them to be *correct* or *good* or *appropriate*?





| **Specific Properties**<br>*Does an architecture explain how a system has these?* | Language | Images | Maps |
|---|---|---|---|
| Systematicity | ✔ | | |
| **Logical operators** (and, or, not, implies...) | ✔ | | |
| Categorization | (✔) | | |
| Partial information | ✔ | | |
| More detail | | | |

*What it results in; not how it happens!*

## More Detail — Language

ob Ford made suggestive remarks about his driving habi
under the influence in Dec. 2011. Nevertheless, the office
nder oath. Drivers at the mayor's office admitted that the
nce that time. Whether the issue will become important
*nd Mail* reported had disappeared and did not return our

Suppose we want to look at those "suggestive remarks" more closely. We are unlikely to learn much by looking at the *textual representation* more closely...





| **Specific Properties** *Does an architecture explain how a system has these?* | Language | Images | Maps |
|---|---|---|---|
| Systematicity | ✔ | | |
| **Logical operators** (and, or, not, implies...) | ✔ | | |
| Categorization | ✔ | | |
| Partial information | ✔ | | |
| More detail | ✘ | | |





| **Specific Properties** *Does an architecture explain how a system has these?* | Language | Images | Maps |
|---|---|---|---|
| Systematicity | ✔ | ✘ | |
| **Logical operators** (and, or, not, implies...) | ✔ | ✘ | |
| Categorization | ✔ | ✘ | |
| Partial information | ✔ | ≈ | |
| More detail | ✘ | | |





## More Detail — Images

## More Detail — Images





| Specific Properties<br>*Does an architecture explain how a system has these?* | Language | Images | Maps |
|---|---|---|---|
| Systematicity | ✔ | ✘ | |
| **Logical operators** (and, or, not, implies...) | ✔ | ✘ | |
| Categorization | ✔ | ✘ | |
| Partial information | ✔ | ≈ | |
| More detail | ✘ | ✔ | |





| Specific Properties<br>*Does an architecture explain how a system has these?* | Language | Images | Maps |
|---|---|---|---|
| Systematicity | ✔ | ✘ | ? |
| **Logical operators** (and, or, not, implies...) | ✔ | ✘ | ? |
| Categorization | ✔ | ✘ | ? |
| Partial information | ✔ | ≈ | ✔ |
| More detail | ✘ | ✔ | ? |

# Networks and Machine Learning ("Brain-Style Computing")





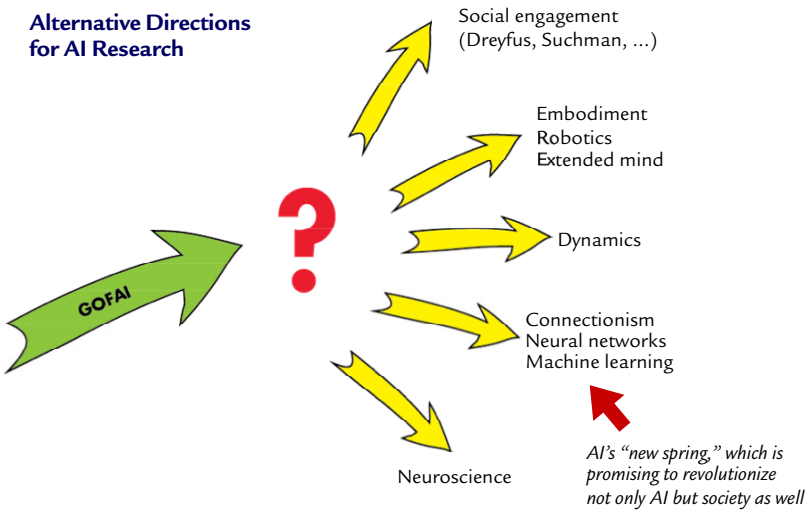
## The Failure of GOFAI

1. As we've seen, there are many deep and powerful lessons and insights in both:
   a) The GOFAI model in particular—such as its ability to deal with systematicity, productivity, and compositionality, through its legacy basis on logic; and
   b) The more general model of a mechanically effective system whose structures and processes are subject to the normative constraints deriving from its semantical (representational) relation to the world.
2. Yet, since the 1980s, GOFAI itself is widely thought to have **failed**
3. So what have AI and cognitive science being **investigating instead**?





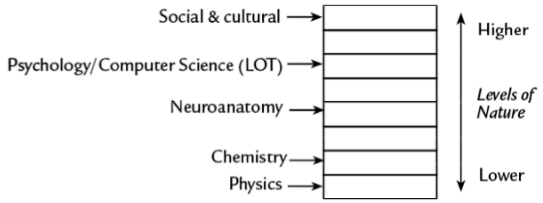
## Alternative Directions for AI Research



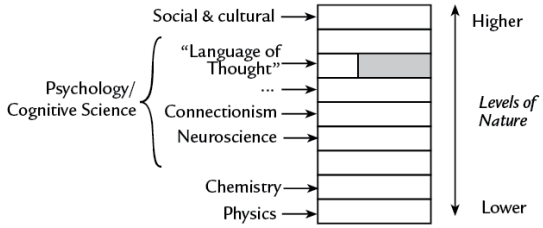





## The "Hierarchy of Nature" — Traditional View (~1970s)



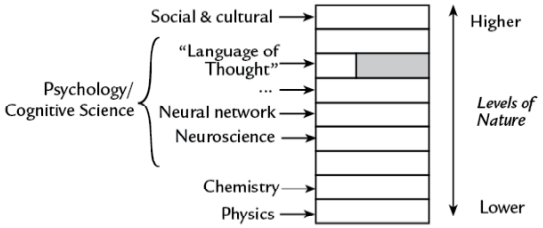
## The "Hierarchy of Nature" — Recent View (~2010)








## The "Hierarchy of Nature" — Contemporary View (~2017)








## Alternative I: Connectionism/neural networks ("brain-style" computing)

Based on 3 (more or less explicit) claims

1. Only **some** (perhaps very **little**) human cognition is explicable in terms of anything *language-like*
2. There is a **lower level** that is much more **general**
   a) I.e., explanations framed in terms of this lower level will be able to account for a *much larger percentage* (perhaps even *all*) of human cognition
3. That (small) fraction that *does* require adversion to explicit (language-like) representation will be explained in terms of how it is **implemented on top** of a connectionist or other lower-level brain-style account





## Observations

1. Though *inspired* by the brain, connectionist or neural-network architectures still operate at a **level of abstraction** *substantially above actual neuroscience*
2. Though neural-network architectures (and machine learning systems) are supported by huge numbers of enthusiasts today (just as GOFAI was, 40 years ago!), the most important recognition to come out of it may be that the human psyche *may not have a unique explanation at any one single level of analysis*
3. Rather, **multiple different levels** of analysis may figure simultaneously in an account of human cognition
4. This raises interesting questions about the forms of intelligence we should expect to see in (future) artificial systems

**Issue:** *As we study machine learning and network systems, keep this question at the front of your mind: what is this architecture good at explaining?*

**Caveat:** *Don't believe claims that connectionist, network, or (machine-learning) neural architectures are not computational!*

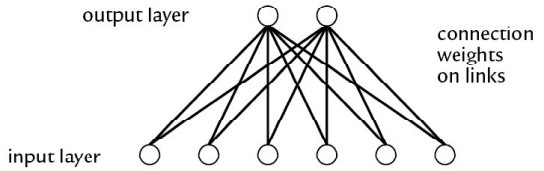
## Properties of the brain

1. **Very slow** (order of 10–100 milliseconds per "operation")
2. **Massively parallel** (order 10 neurons, 10 connections)
3. Feldman's "100 step rule":
   - To figure out how the brain solves a problem, you have to figure out how it can do it in no more than **100 serial steps** of a **massively parallel architecture**.





## Simple (two-layer) "Feed-Forward" network

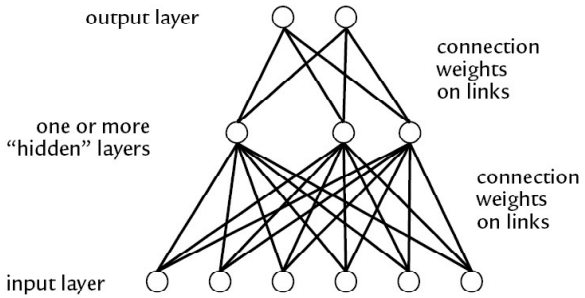



1. At least two **layers** ("**input**," "**output**", and possibly some "**hidden**" layers)
2. **Connection weights** on the links (usually modelled with real numbers between 0.0 and 1.0)
3. Nodes that **combine** the weights on the (incoming links)
   a) Assumed to be a simple arithmetic calculation
   b) Often: **summation** (or a sigmoid function—Σ)
4. Often: only output a signal if value is greater than some **threshold**





## Three (or more) layer networks



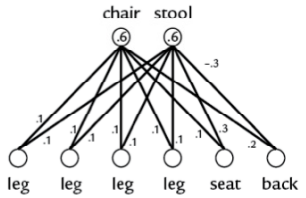





## Chair/stool recognizer

**Activation rule**

1. Simple summation
2. Threshold
   a) 0 if input < 0.6
   b) 1 if input ≥ 0.6



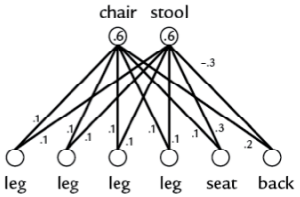

| | Leg | Leg | Leg | Leg | Seat | Back |
|---|---|---|---|---|---|---|
| Chair | .1 | .1 | .1 | .1 | .1 | .2 |
| Stool | .1 | .1 | .1 | .1 | .3 | -.3 |

## Chair/stool recognizer (cont'd)

0.7/0.4 ⟹ chair

0.4/0.6 ⟹ stool

0.7/0.4 ⟹ chair




0.5/0.7 ⟹ stool

0.6/0.4 ⟹ chair



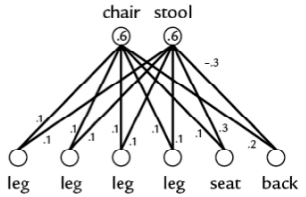


## Chair/stool recognizer (cont'd)

Unfortunately, this recognizer isn't very sophisticated—lots of legitimate chairs and stools are misclassified by it...

0.4/0.1 ⟹ *nothing*

0.5/0.2 ⟹ *nothing*




0.2/0.4 ⟹ *nothing*

0.3/0.5 ⟹ *nothing*





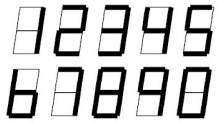
## Notes on the chair/stool recognizer

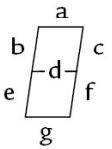
1. Inputs: *unrealistically "high level"* (recognition recurses!)
2. Weights: relatively small numbers—not too different
3. Features are *too abstract and unrelated* (nothing "holistic" about its recognition)





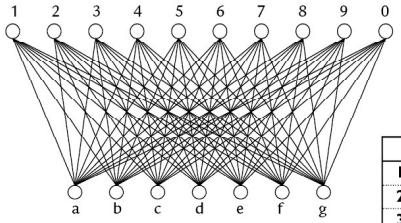
## Digit recognizer






| | a | b | c | d | e | f | g |
|---|---|---|---|---|---|---|---|
| 1 | | | ✔ | | | ✔ | |
| 2 | ✔ | | ✔ | ✔ | ✔ | | ✔ |
| 3 | ✔ | | ✔ | ✔ | | ✔ | ✔ |
| 4 | | ✔ | ✔ | ✔ | | ✔ | |
| 5 | ✔ | ✔ | | ✔ | | ✔ | ✔ |
| 6 | | ✔ | | ✔ | ✔ | ✔ | ✔ |
| 7 | ✔ | | ✔ | | | ✔ | |
| 8 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| 9 | ✔ | ✔ | ✔ | ✔ | | ✔ | |
| 0 | ✔ | ✔ | ✔ | | ✔ | ✔ | ✔ |

## Digit recognizer (cont'd)




| | a | b | c | d | e | f | g |
|---|---|---|---|---|---|---|---|
| 1 | .3 | -.3 | .3 | -.3 | -.3 | .3 | -.3 |
| 2 | .1 | -.1 | .1 | .1 | .1 | -.1 | .1 |
| 3 | .1 | -.1 | .1 | .1 | -.1 | .1 | .1 |
| 4 | -.3 | .13 | .13 | .13 | -.3 | .13 | -.3 |
| 5 | .1 | .1 | -.1 | .1 | -.1 | .1 | .1 |
| 6 | -.1 | .1 | -.1 | .1 | .1 | .1 | .1 |
| 7 | .2 | -.2 | .2 | -.2 | -.2 | .2 | -.2 |
| 8 | .07 | .07 | .07 | .07 | .07 | .07 | .07 |
| 9 | .1 | .1 | .1 | .1 | -.1 | .1 | -.1 |
| 0 | .9 | .9 | .9 | -.9 | .9 | .9 | .9 |





## Notes on "digit" (re)cognizer

**Problems**

1. Inputs
   a) More plausible (than what?)—but still not *biological*
   b) More local (than a "table from left field")—but still not "*points*"
2. Weights
   a) Relatively small numbers, not too different (again)
3. Setup
   a) How are the weights established?
   b) How can we learn?

**Solutions**

4. All these issues can be addressed!
   a) Can make the models more biologically plausible
   b) For at least some recognition tasks, can go to point inputs
   c) Train the networks—build up the weights automatically





## Learning

1. The most important fact about these networks: they can **learn** !
2. There are numerous ways to **train** them
3. The simplest is called **back propagation**
   a) Give it an input
   b) Compare what the output **is** to what it **should be**
   c) Propagate "the difference" (or something based on it) back down to the connections. E.g.:
      i. If activation state is *too high*, *decrease* the weights of positive connections (and *increase* the magnitude of the connections that contributed negatively).
      ii. If the activation state is *too low*, do the opposite ...
4. A strategy currently gaining a huge amount of attention is **predictive coding**: predict what you think is going to happen, compare the "input" (what *is* happening) against that, and propagate upwards only the error—i.e., the difference between what you expected and what you actually encountered.
5. It is **stunning** to see what these systems can learn





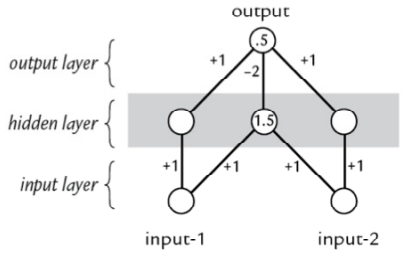
## Dynamics

1. Leads to two forms of **dynamics**
   a) **High-speed dynamics**: produce an output, given an input
   b) **Low-speed dynamics**: adjust the connection weights, to learn
2. On current systems, it can take **many thousands** (sometimes **millions!**) of "training runs" before the connection weights settle down.

## Simple Examples

1. An example of a route follower
   — **http://www.youtube.com/watch?v=0Str0Rdkxxo**
2. Simple handwritten digit recognizer
   — **https://www.youtube.com/watch?v=ocB8uDYXtt0**
   — **https://algorithmia.com/demo/handwriting**
3. Handwriting *generator*
   — **www.cs.toronto.edu/~graves/handwriting.cgi**
4. Ball balancer
   — **https://www.youtube.com/watch?v=Bk2oDaYeRiQ**
5. A famous early "speaker" of English words (NetTalk)
   — **http://www.youtube.com/watch?v=gakJlr3GecE**
6. More examples
   — **http://deeplearning.net/demos/**

## More complex types of network …

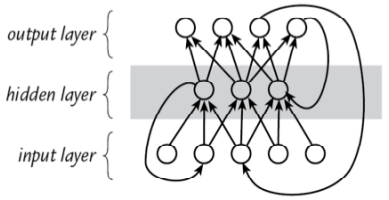



1. "Exclusive or" — requires a **hidden layer**
2. One of the challenges with hidden layers was that, when these networks were first being explored, it wasn't clear what algorithms could be used to **train** the hidden layers
3. The recent upsurge in interest in (and power) of deep learning systems is that researchers have **figured out ways to train hidden layers**. Networks are now being trained with up to dozens of hidden layers…





## A recurrent network




1. **Recurrent** networks are networks where the outputs of at least some of the nodes are fed back and used as inputs to other nodes.
2. Among there things, they are used for processing temporal patterns, and exhibiting **dynamic behaviour**.
3. But recurrent networks present their own **training** challenges (as well as issues of **network stability**!)





## Temporality and Dynamics

1. Somehow, we remember *dynamic patterns and rhythms*—like songs—in passive form
   - If someone asks you to sing a song you know well, you don't have to wait till the first line "comes around"!
2. It is likely that we use cascades of oscillators—**controlled instability** of a sort
3. But the nature of such networks, and how to train them, remains (as far as I know!) well beyond what we current know how to do.





## Formalization

These networks can all be "formalized" (i.e., a theory of them can be developed that characterizes these networks in mathematical terms):

Units: $1 \text{–} N$

Activation state: $\sigma_i(t)$

Connection weights: $w_{ij}$ (weight of link $i \rightarrow j$)

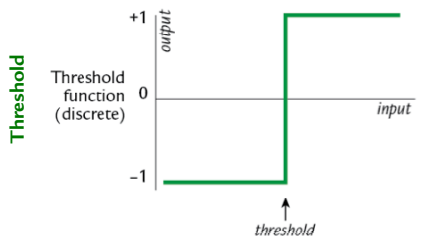
Activation rule: $\sigma_j(t+1) = f_j(w_{ij}\sigma_i(t))$

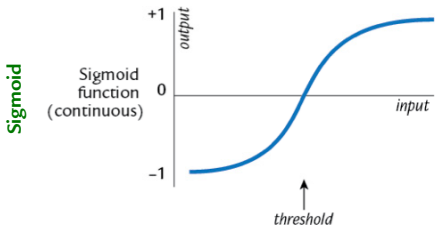
Summation rule: $\sigma_j(t+1) = \sum_i(w_{ij}\sigma_i(t))$

(Threshold) Output rule: $\beta_i(t) = 0$ if $\sigma_i(t) < \xi$, $1$ otherwise

*Lots of variations for continuous activiation values, time, etc. E.g.:*

(Sigmoid) Output rule: $\beta_i(t) = \frac{1}{1 + e^{-\alpha_i(t)}}$








## Instead of formalizing (in this class), we will ask two simple questions

**Q1** What are these networks **good at**?

a. And what are they *not* good at?

**Q2** What are these networks **doing**?

a. Are they *representing*?

b. Are they *inferring* (thinking)?



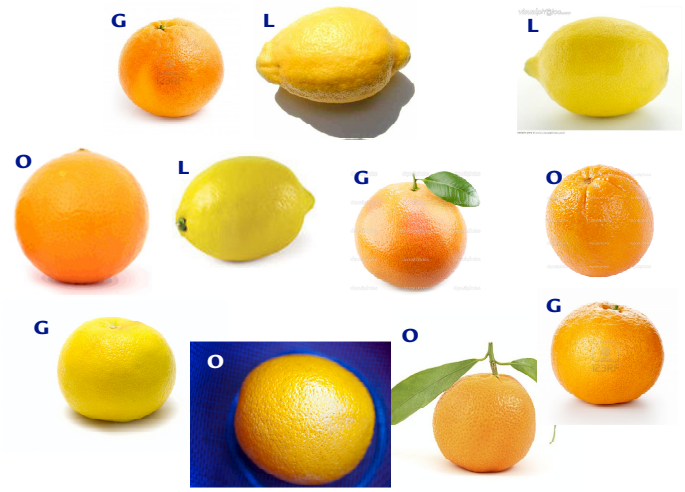


## Q1 — What are these networks good at?

1. **Pattern recognition** / best-match search
   a) E.g., chicken-sexing
   b) E.g., oranges vs. grapefruits (vs. clouds!)
   c) E.g., object-tracking (for camera focusing)







## Q1 — What are these networks good at?

1. **Pattern recognition** / best-match search
   a) E.g., chicken-sexing
   b) E.g., oranges vs. grapefruits (vs. clouds!)
   c) E.g., object-tracking (for camera focusing)





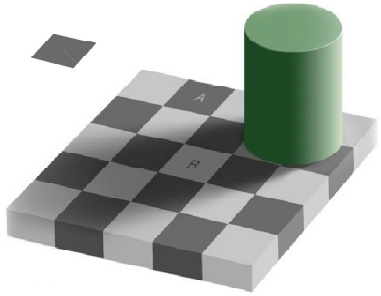
## Q1 — What are these networks good at (cont'd)?

1. **Pattern recognition** / best-match search
   a) E.g., chicken-sexing
   b) E.g., oranges vs. grapefruits (vs. clouds!)
   c) E.g., object-tracking (for camera focusing)
2. **Constraint-satisfaction problems**
   a) 3D constraints on (from) 2D projections





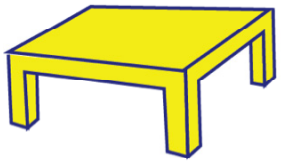
1. Use '**diamond**' to refer to the rhomboid-shaped regions in the picture, and '**tile**' to refer to the square regions on the chess board that is represented by the picture.
2. Surprisingly, the *diamond* labeled 'A' is exactly the same colour as the *diamond* labeled 'B'.
3. Some would call this an *optical illusion*—but I believe that label is mistaken!
4. After all, the *tile* labeled 'A' is a different colour than the *tile* labeled 'B'.
5. It is the (3D) *tiles*, not the (2D) *diamonds*, that we "see" when we parse/interpret the picture.
6. So our perceptual systems are **doing the right thing** (even if the 2D facts that *tell us that truth* about the 3D world might not be what we would naively expect).

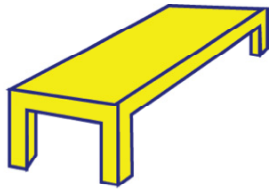







1. Similarly, the 2D quadrilaterals used in the pictures below to indicate the tops of two tables are identical in shape! (shown in blue, to the right)
2. Many would call this an optical illusion, too—but I disagree, since what our visual attention is directed towards is the **3D table depicted**, not the *2D depiction of that table*.
3. And the tables do *not* have similarly shaped *tops*.





## Q1 — What are these networks good at (cont'd)?

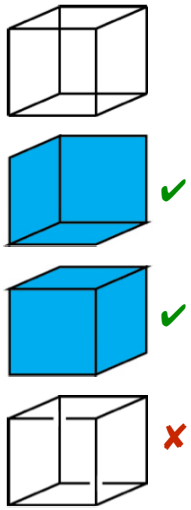
1. **Pattern recognition** / best-match search
   a) E.g., chicken-sexing
   b) E.g., oranges vs. grapefruits (vs. clouds!)
   c) E.g., object-tracking (for camera focusing)
2. **Constraint-satisfaction problems**
   a) 3D constraints on (from) 2D projections





## Q1 — What are these networks good at (cont'd)?

1. **Pattern recognition** / best-match search
   a) E.g., chicken-sexing
   b) E.g., oranges vs. grapefruits (vs. clouds!)
   c) E.g., object-tracking (for camera focusing)
2. **Constraint-satisfaction problems**
   a) 3D constraints on (from) 2D projections
   b) Necker cubes
   c) Content-addressable memory
   d) Similarity-based generalization







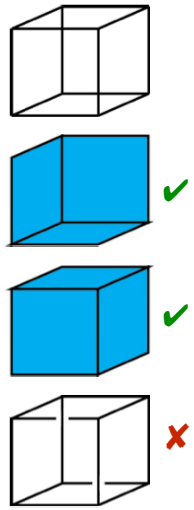






## Q1 — What are these networks good at (cont'd)?

1. **Pattern recognition** / best-match search
   a) E.g., chicken-sexing
   b) E.g., oranges vs. grapefruits (vs. clouds!)
   c) E.g., object-tracking (for camera focusing)
2. **Constraint-satisfaction problems**
   a) 3D constraints on (from) 2D projections
   b) Necker cubes
   c) Content-addressable memory
   d) Similarity-based generalization
   e) Graceful degradation (in the presence of noise, interference, damage, and overload)








## Q2 — What are these networks doing?

We'll talk about that in the next lecture (C·03)

## Conclusion: III · A — Connectionism & Neural Networks ("Brain-Style Computing")

1. Last Thursday (Lecture 07·b) we introduced our first "alternative" model of the mind—**connectionist/neural-network ("brain-style") architectures**—and explored a bit about what such systems were *good at* (pattern-recognition, constraint satisfaction, etc.)
2. Today, we will conclude our discussion of such "brain-style" architectures by focusing on *what they are doing*—not in the detail sense of how they work in detail, but asking questions about the nature of mind they imply (e.g., are they representational?).





## Are connectionist/neural-network models *representational*?

*That is: do they fit into the "***general model***" we talked about at the end of Part II?*

1. Many proponents of neural network models are vociferously *anti-representational!*
2. But the actual answer isn't that clear, for two reasons:
   a) First, at the network level itself, such systems can be understood as **representing a large number of domain micro-features**
      i. Even the inputs are likely representational in *some sense* (e.g., representations of the intensity of light hitting a sensor element?)
      ii. More seriously, if there is any coherence or regularity to the patterns of activation or connection strengths that enable them—e.g., to recognize faces or shapes or constraints—then it would seem that those patterns of activation or connection strength represent aspects or features of the shapes that they ultimately recognize.
      iii. It may be that the representations aren't **explicit**, in the sense of being "objects" that a separate locus of activity can *manipulate*, in the way that CPUs and other "inner processes" manipulate data structures in classical computational architectures (cf. Lecture 07·a).
   b) Also, aren't the patterns of activation **normatively governed** by relations to the external world or task domain (i.e., connected with blue arrows of some sort)?





## Are connectionist/neural-network models *representational* (cont)?

   c) Various machine learning experts—including Geoff Hinton—believe that these systems do develop representations
   d) Plus, independently of the basic vectors of connection strength, is also possible that high-level representations can be **implemented upon**, or can **emerge from**, an underlying neural-network base.
      i. This possibility would lead to lots of questions
      ii. E.g.: Would the emergent high-level representational capacities inherit the properties of graceful degradation under *noise*, *damage*, *overload*, etc. that we saw to be characteristic of the lower levels?
3. In sum, there isn't general theoretical agreement in the field—in part because

**We don't (yet) have a generally accepted theory of what it is to be representational**

4. For now, therefore, it is probably most productive:
   a) **Not** to think of neural networks as *non-representational*
   b) Instead to think of them as **a different kind of (representational) architecture**
5. We will want to keep an eye on this issue of representation through the next several alternative architectures. Towards the end of the course I will propose a better understanding of when—and why—systems are, and are not, representational.





## A famous debate (slugfest)

Related to the question of whether connectionist/neural networks are representational is a question that has generated a huge (and famous) debate:

Q: Are connectionist systems (neural networks) compositional, systematic, and productive?

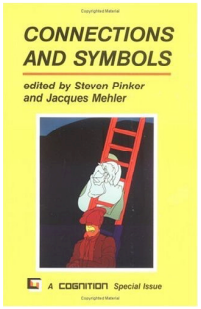
**Fodor & Pylyshyn**: *Either*

a) Networks **cannot** exhibit compositionality, systematicity, and productivity—in which case they *aren't even candidates to be an* (or the) *architecture of mind*; or
b) They **can** exhibit compositionality, systematicity, and productivity, in which case they merely **implement** a language of thought (LOT)—in which case the LOT explanation is the important one, and the fact that they are *networks* is *psychologically irrelevant* (no relevance to mind)
   – Cf. criterion **C2**, on slide 3 of Lecture 07·a

**Smolensky**:

a) These networks **can** exhibit compositionality, systematicity, and productivity—but
b) They do **not** do this by implementing a language of thought (LOT)—and therefore they are *not* merely "implementations" of a classical architecture
c) Hence networks are a new (and good) model of the architecture of mind, a **genuine alternative to LOT**

## A famous debate (slugfest) … cont'd

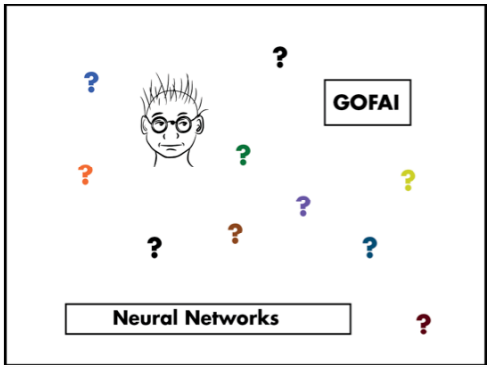



1. Some of the papers in this debate are collected in a book edited by Steven Pinker and Jacques Mehler entitled "*Connections and Symbols*" (1988)
2. The debate has died down, but it isn't really **resolved**
3. My own sense: three things are at play:
   a) We don't understand '*implementation*' well enough to be able to settle the issue once and for all
   b) Each architecture suggests a powerful explanation of *some* aspects of human cognition
   c) **Neither** architecture, on its own, can account for *everything* of importance about the human mind!
4. We should therefore view the two architectural proposals as either:
   a) Potential ingredients in a mind (*parts* of how we work?); or
   b) Explorations of two points in a large (and as yet most unexplored) design space of possible architectures.





## How much of the "design space" have we explored yet?








## Issues about neural-network architectures

1. **Opacity** of explanation
   a) What do *we learn* when a network accomplishes a task?
   b) Is the *architecture* (connectivity) of the nodes what matters
      i. Or the resulting *connection weights*?
      ii. Or the *learning algorithm*?
      iii. Or all three?
      iv. If all three, what is their relative importance/priority?
2. **Generality:** If *one* network can be trained to accomplish the task, how many *other* networks (with more or fewer nodes, different connectivity, etc.) could do the same? Even if the nodes and connectivity are the same, how many other configurations of connection weights would accomplish the task? What is it about a successful network that matters to its success?
2. **Ineffability** of internal states
   a) What has *the network learned*, when the training is done?
   b) What does it know?
   c) E.g.: does it think that eyebrows are important (in face recognition)? Or skin colour (for discriminating oranges from grapefruits)?

*We don't really have good answers to these questions*





## Issues about neural-network architectures (cont'd)

4. **Emergence**
   a) It is common to hear that intelligence overall, and many characteristics of intelligence (including representation, systematicity and productivity, etc.), are **emergent** properties of neural networks.
   b) Emergence is one of the trendiest—but most difficult to understand—notions in contemporary cognitive science (and many other fields)
   c) For example: is emergence an *epistemological* or *ontological* notion?
      i. **Epistemological:** Are we just *surprised* that some behaviour/result arises out of a base system, even though in fact it is completely determined by it (and would be entirely predictable, if only we were smarter)?
      ii. **Ontological:** Or is that the behaviour/result in question is actually not "reducible to" the ingredients out of which it stems—somehow not a result of their individual properties and relationships?
      iii. E.g.: Termite mounds, birds' and insects' "swarming," etc.—how are these things to be explained?
   d) (Note that the PhD dissertation of Joel Walmsley—the author of our *Mind and Machine* textbook— was an argument that emergence is *only* an epistemological notion—that "ontological emergence" is not a sensible concept!)

## Issues about neural-network architectures (cont'd)

5. **State spaces**
   a) The activation states on *n* nodes can be taken to be a point in an *n*-dimensional state space
   b) Similarly, the weights on the connections between these nodes can be taken as a point in an $n^2$-dimensional state space
   c) To what extent is this kind of state-space characterization useful?
   d) What properties, that are illuminating about the mind, derive from the structure of the state space; what have to do with the particular shape of (or trajectory through) a state space?
   e) We will see some insights in this direction when we look at **dynamical systems** (the next "alternative architecture"), but many of these questions remain open research problems.





## Issues about neural-network architectures (cont'd)

6. **Ethics**
   a) The fact that we can't necessarily understand what they've learned, or how some things they "know" interact with other things they "know," leads to complex ethical issues
   b) These facts interact with the fact that they are typically trained on huge amounts of social data—which can reflect biases and prejudices embedded in the cultural milieu
      i. Cf. revelations that a search for "unprofessional hair" on Google images returned large numbers of pictures of black women, as opposed to "professional hard" producing far more pictures of white women.*
      ii. Cf. firestorm of protest when Google images classified black people as "gorillas."†
   c) Ethical questions are likely to come more and more into focus, as these systems are developed and deployed in society.

† http://www.cbc.ca/news/trending/google-photos-black-people-gorillas-1.3135754

* http://www.telegraph.co.uk/technology/2016/04/08/google-under-fire-over-racist-image-search-results-for-unprofess/





## Comparison

The best way (i) to understand the relationship between neural-network architectures and logic/GOFAI based systems, and (ii) at least to begin to understand the wider design space of possible architectures, is in terms of the following point-by-point comparison:

**A. Neural Networks** — involve
1. *Shallow* (few step) inference
2. On *massive amounts* of data
3. Involving *very large numbers* of
4. *Weakly correlated* variables

**B. Logic (GOFAI) systems** — involve
1. *Deep* (many step) inference
2. On *modest amounts* of information
3. Involving a *small number* of
4. *Strongly correlated* variables

**$1M questions:**

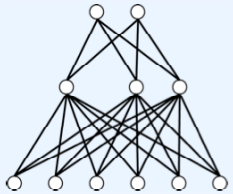
a) Does mind involve some *combination* of those two architectures?
b) Would a combination involve a logic/GOFAI (or similar) system *implemented on top of part* of a neutral network?
c) Are there intermediate positions between neural networks and logic/GOFAI systems? Is it a continuum?
d) Are there ways in which neural networks and logic/GOFAI systems are similar, which are *not* be shared by another plausible or powerful mental architecture?

Cf. Fodor and Smolensky!





## Summary of Connectionism / Neural Networks



**A. What are they? Brain-style computing**
1. Modelled on (but more abstract than) the brain;
2. Slow connections among massive numbers of parallel parts (remember Feldman's rule)
3. Different from GOFAI (but you can *implement* GOFAI on top of a neural network)

**B. Performance**
1. (Stunningly) good at: *pattern matching*, *classification*, and *constraint satisfaction*
2. May also be able encode simple implications in the distributed weights*
3. No evident ability to deal with "not," "or", or other complex logical relations
4. No evident ability to deal with deep or deliberative reasoning

**C. Mechanism:**

| Networks | GOFAI |
|---|---|
| 1. Shallow (few step) inference | 1. Deep (many step) inference |
| 2. On massive amounts of data | 2. On modest amounts of information |
| 3. Involving very large numbers of | 3. Involving a small number of |
| 4. Weakly correlated variables | 4. Strongly correlated variables |

E.g., that small dogs are likely to have higher barks than large ones, that cats and dogs are more similar to each other than either is to a tree or to a French horn, etc.

## Summary of Connectionism / Neural Networks (cont'd)

**D. Trainin**

1. Tremendously important: these networks can be *trained*
2. At least presently, can take *large amounts of time* being trained on *very large data sets*
3. In a sense: successful as a result of "big data" and "high performance computing"

**E. Issues**

1. Theoretically, it isn't clear *how much we do—or can!—understand them*
2. Therefore hard to know what we can—and cannot—trust them with
3. Serious ethical issues arise when they are trained on human data sets (e.g. Twitter)
4. Questions being raise about job displacements, as these systems "take over"

**F. State of play**

1. Dramatic recent successes: Deep Mind's AlphaGo program defeating world "Go" champion Lee Sedol, driverless cars, face and image recognition (surveillance), etc.
2. Neural network systems based on deep-learning will increasingly permeate our lives
3. It will be vital, in the next 10 years and more, to know what we want to have these systems do, and what we want to reserve for humans.

**G. Bottom line**

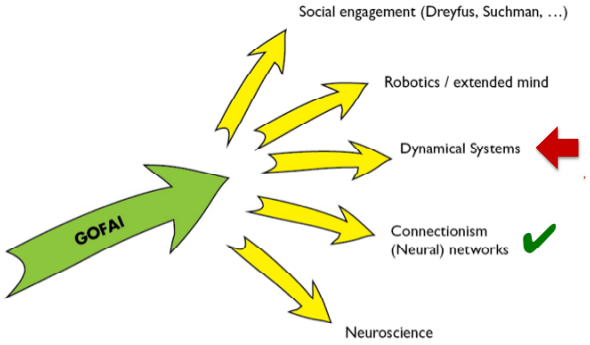
1. For what they do .... ... **extraordinarily impressive**
2. Are they part of the architecture of mind? ... **very likely**
3. Are they *the* (complete) architecture of mind ... **very unlikely**





## Part III · B

# Dynamical Systems

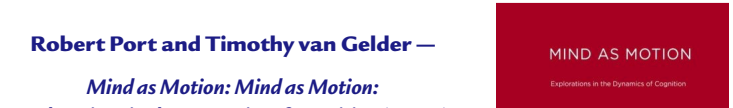







**Robert Port and Timothy van Gelder —**

***Mind as Motion: Mind as Motion:***
***Explorations in the Dynamics of Cognition* (1995)**




Bob Port

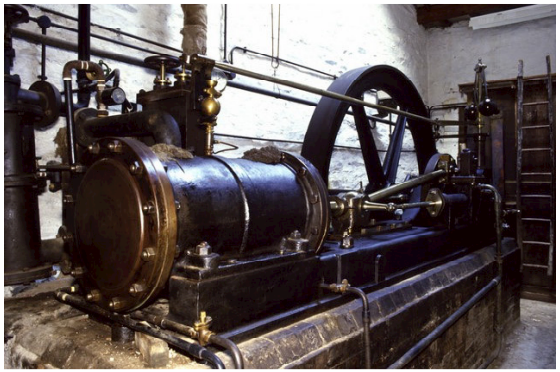
Tim van Gelder

Read: Timothy van Gelder, "The Dynamical Hypothesis in Cognitive Science" (on Blackboard)





## Task: control the speed of a steam engine



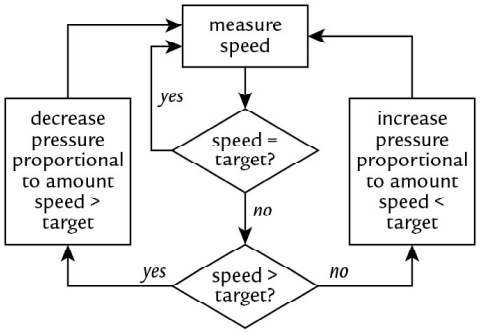




## Computational version — description

1. **Begin**
   a) Measure the speed of the flywheel
   b) Compare the actual speed against the desired speed
2. **If there is no discrepancy, return to step 1**
3. **If there is a discrepancy**
   a) Measure the current steam pressure
   b) Calculate the desired alteration in steam pressure
   c) Calculate the necessary throttle valve adjustment
   d) Make the throttle valve adjustment
   e) Return to step 1

## Computational version — flowchart








## Computational version — code

```
begin proc control(target::fix)
     local (speed::fix, pressure::fix)
   1 speed := measure-motor-speed()
   2 pressure := measure-steam-pressure()
   3 if speed = target go to 1
   4 if speed < target go to 7
   5 set-steam-pressure(min(0,pressure*(1-((speed-target)/target))))
   6 go to 1
   7 set-steam-pressure(max(0,pressure*(1+((target-speed)/target))))
   8 go to 1
end
```

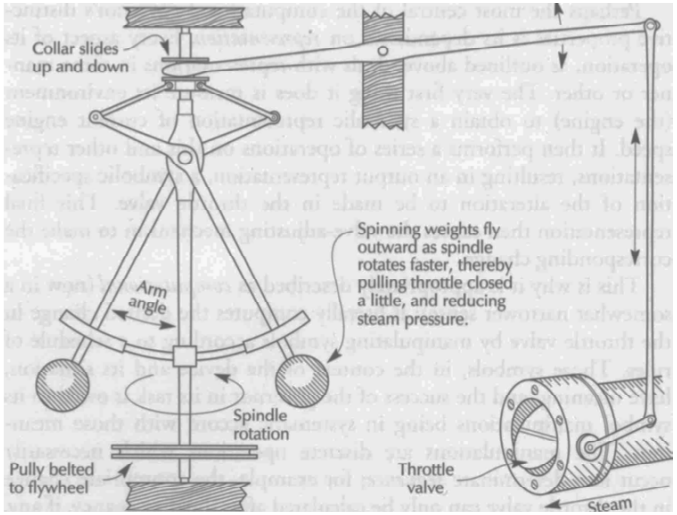
*NB: this is absurd code ;-)*
*much more reasonable would be this:*

```
(while (true)
       (set-steam-pressure(min(0,
          (measure-steam-pressure() *
             (2 - (measure-motor-speed()/target)))))
```





## Watt Governor





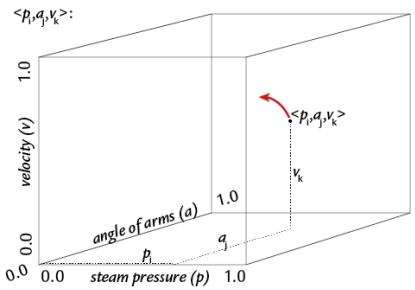



## Watt Governor (cont'd)

1. **Explanations**
   a) http://www.mekanizmalar.com/flyball_governor.html
   b) https://www.youtube.com/watch?v=SiYEtnlZLSs
   c) http://www.mekanizmalar.com ⟸ *an interesting site!*
2. **Examples**
   a) http://www.youtube.com/watch?v=R6rgZ-7Y3t8
   b) http://www.youtube.com/watch?v=3x3Mo6_8zGc
   c) http://www.youtube.com/watch?v=4GZj3IzXmsE
   d) http://www.youtube.com/watch?v=Nr9UtEhyvfk ⟸ *Papplewick*

## State Spaces

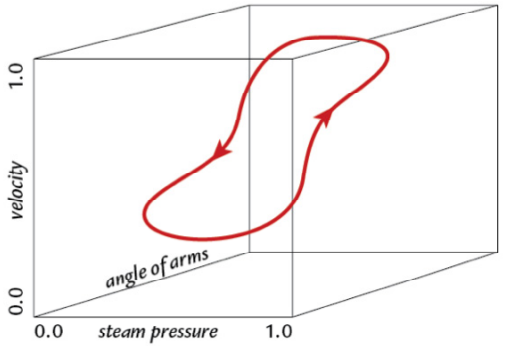
1. The state of a dynamical system can be plotted as a *point in a state space*, where
   a) The "space" is a n-dimensional space of numbers, for each of the n parameters or variables that can vary as the overall system progresses in time (in the case of the governor: the velocity of rotation, the angle of the arms, and the resulting pressure of the steam.
   b) The *progress* of the system can be plotted as a *trajectory* through the state space.







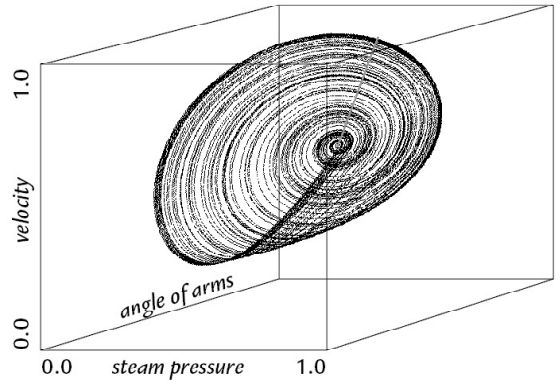

## Governor State Space — "Hunting" (not so good)







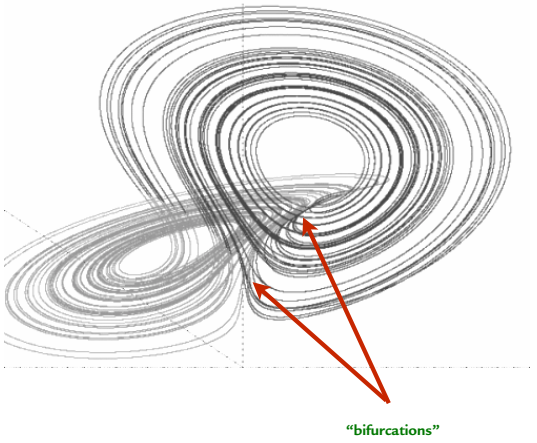

## Governor State Space — Attractor (more what we want)



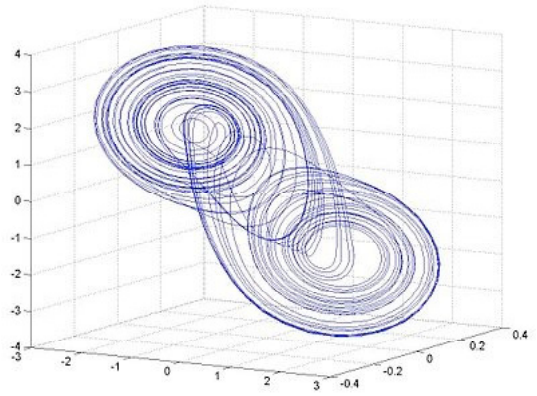





## More complex state spaces — A



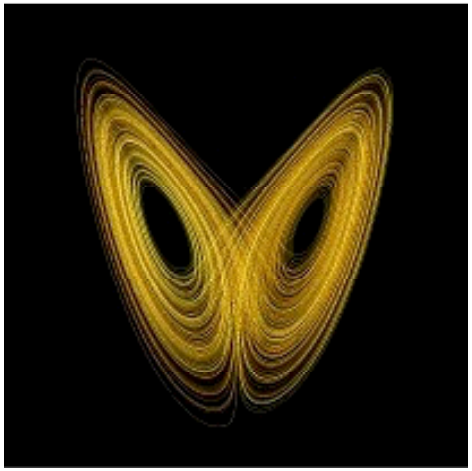
## More complex state spaces — B (two attractors)







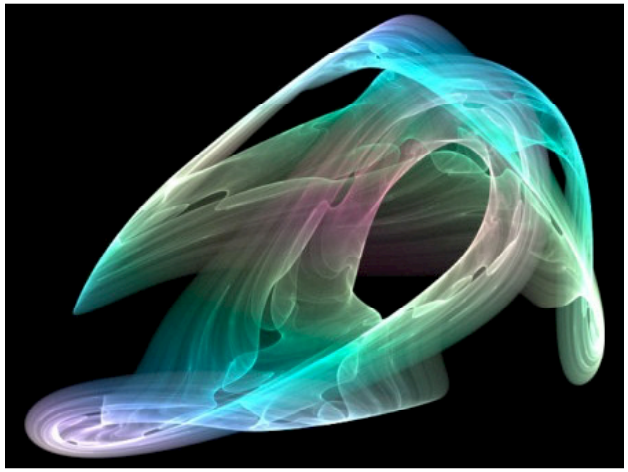

## More complex state spaces — C (two attractors)







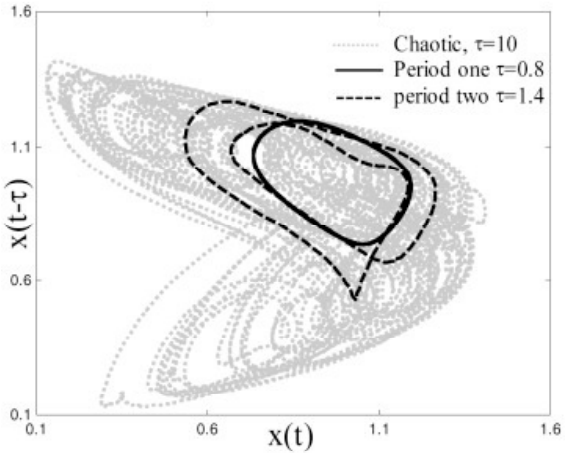
## More complex state spaces — D



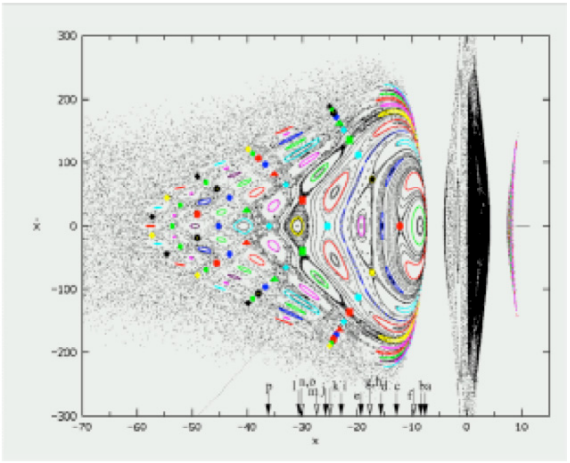




## More complex state spaces — E (chaotic behaviour)



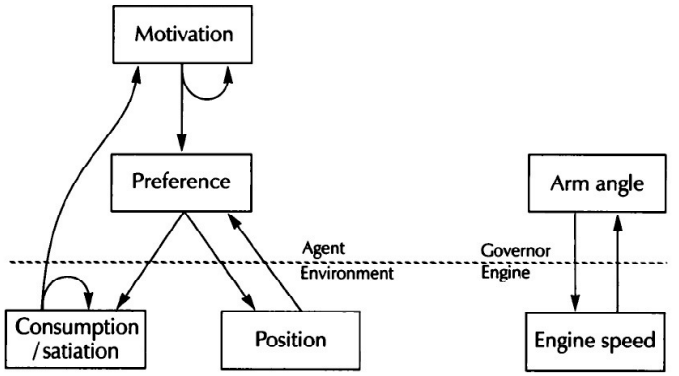
## More complex state spaces — F (chaotic behaviour)







## How does this relate to mind?

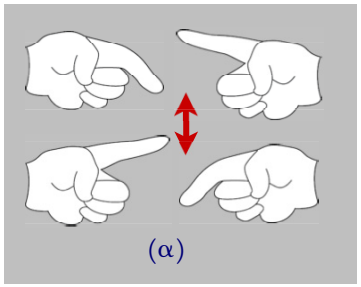


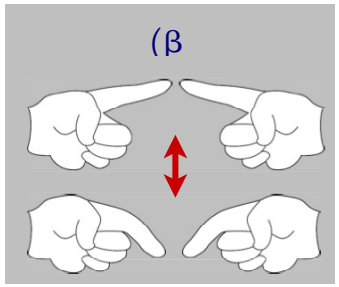





## A famous example

If one starts out wagging one's fingers "out of phase", as in α, and speeds up …




… one inevitably ends up doing it "in phase," as in β





## A famous example (cont'd)

1. This "finger-wagging" behaviour is described by the Haken–Kelso–Bunz (HKB) model of bimanual coordination

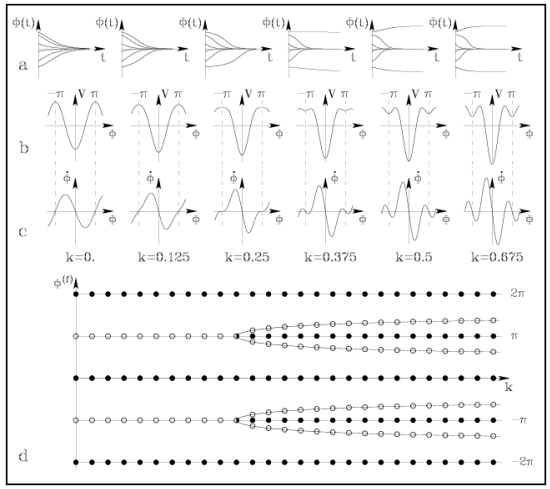
   $V = - a \cos \varphi - b \cos 2\varphi$

   a) φ is the relative phase of the two fingers (0° = in phase; 180° = out of phase)

   NB: φ does not refer to a "part" of any mechanism!

   b) b/a is the frequency

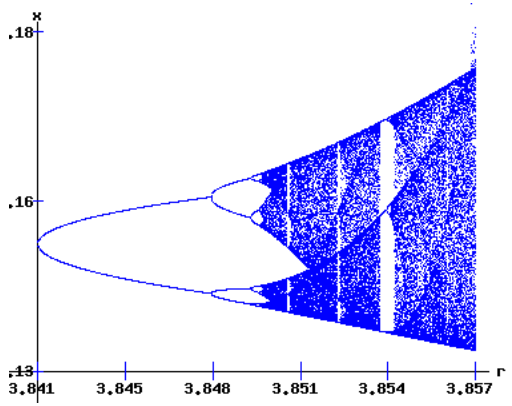
Behaviour of HKB model




Faster ← → Slower












All to be continued …

*On Thursday*

## Part III · B

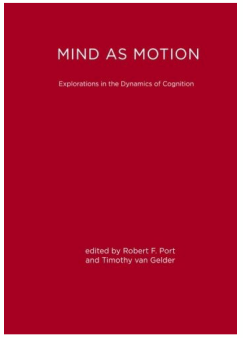
# Dynamical Systems (cont'd)





**Robert Port and Timothy van Gelder —**

***Mind as Motion: Mind as Motion:***
***Explorations in the Dynamics of Cognition* (1995)**




Bob Port

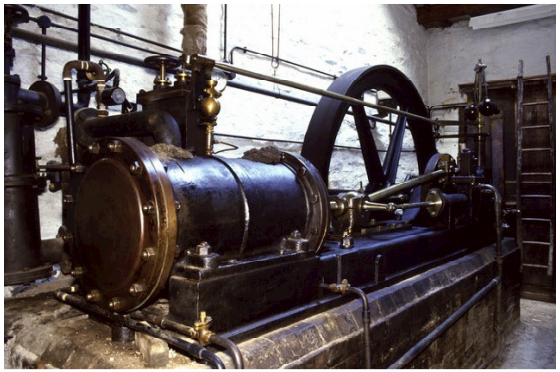
Tim van Gelder

Read: Timothy van Gelder, "The Dynamical Hypothesis in Cognitive Science" (on Blackboard)





## Task: control the speed of a steam engine



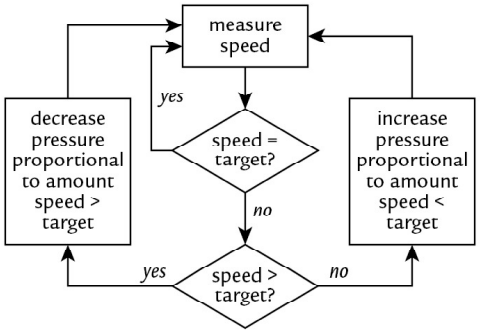




## Computational version — description

1. **Begin**
   a) Measure the speed of the flywheel
   b) Compare the actual speed against the desired speed
2. **If there is no discrepancy, return to step 1**
3. **If there is a discrepancy**
   a) Measure the current steam pressure
   b) Calculate the desired alteration in steam pressure
   c) Calculate the necessary throttle valve adjustment
   d) Make the throttle valve adjustment
   e) Return to step 1

## Computational version — flowchart








## Computational version — code

```
begin proc control(target::fix)
     local (speed::fix, pressure::fix)
   1 speed := measure-motor-speed()
   2 pressure := measure-steam-pressure()
   3 if speed = target go to 1
   4 if speed < target go to 7
   5 set-steam-pressure(min(0,pressure*(1-((speed-target)/target))))
   6 go to 1
   7 set-steam-pressure(max(0,pressure*(1+((target-speed)/target))))
   8 go to 1
end
```

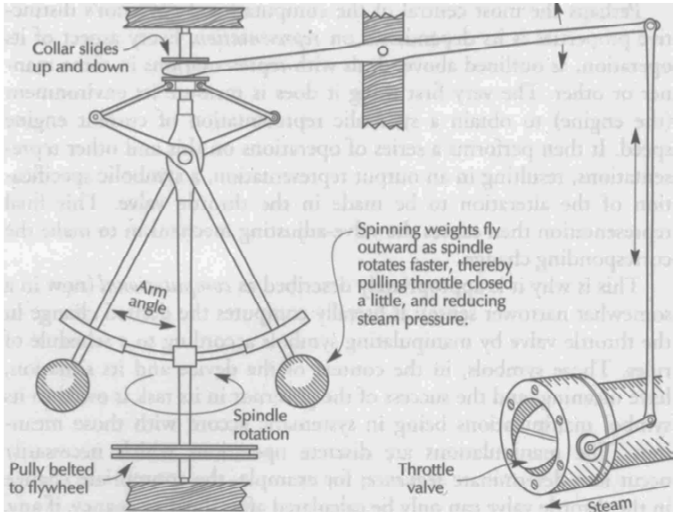
*NB: this is absurd code ;-)*
*much more reasonable would be this:*

```
(while (true)
       (set-steam-pressure(min(0,
          (measure-steam-pressure() *
             (2 - (measure-motor-speed()/target)))))
```





## Watt Governor





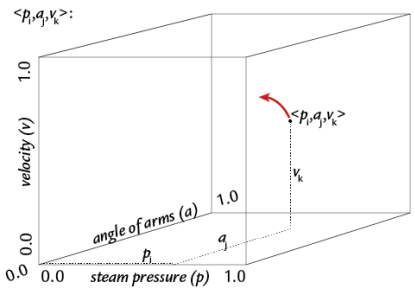



## Watt Governor (cont'd)

1. **Explanations**
   a) http://www.mekanizmalar.com/flyball_governor.html
   b) https://www.youtube.com/watch?v=SiYEtnlZLSs
   c) http://www.mekanizmalar.com ⟸ *an interesting site!*
2. **Examples**
   a) http://www.youtube.com/watch?v=R6rgZ-7Y3t8
   b) http://www.youtube.com/watch?v=3x3Mo6_8zGc
   c) http://www.youtube.com/watch?v=4GZj3IzXmsE
   d) http://www.youtube.com/watch?v=Nr9UtEhyvfk ⟸ *Papplewick*

## State Spaces

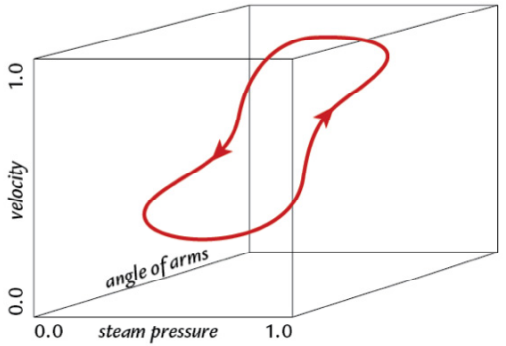
1. The state of a dynamical system can be plotted as a *point in a state space*, where
   a) The "space" is a n-dimensional space of numbers, for each of the n parameters or variables that can vary as the overall system progresses in time (in the case of the governor: the velocity of rotation, the angle of the arms, and the resulting pressure of the steam.
   b) The *progress* of the system can be plotted as a *trajectory* through the state space.







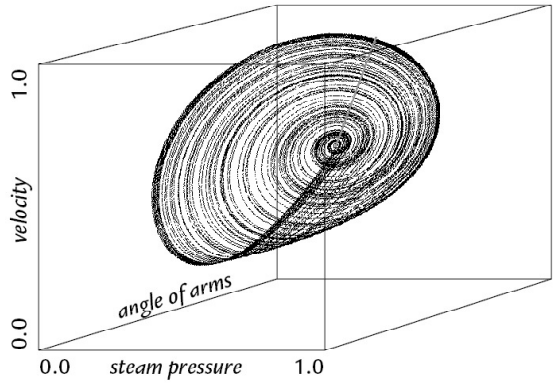

## Governor State Space — "Hunting" (not so good)







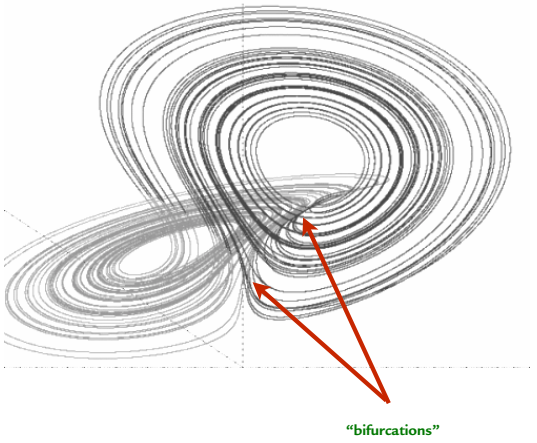

## Governor State Space — Attractor (more what we want)



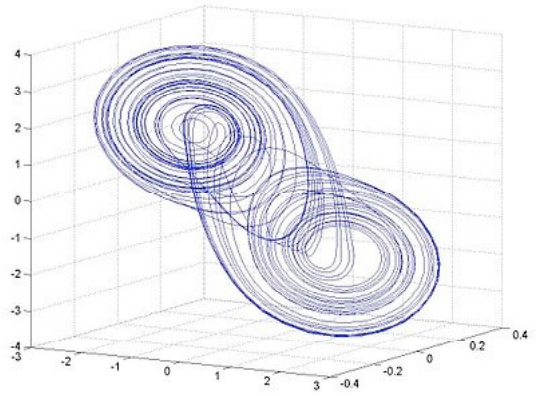





## More complex state spaces — A



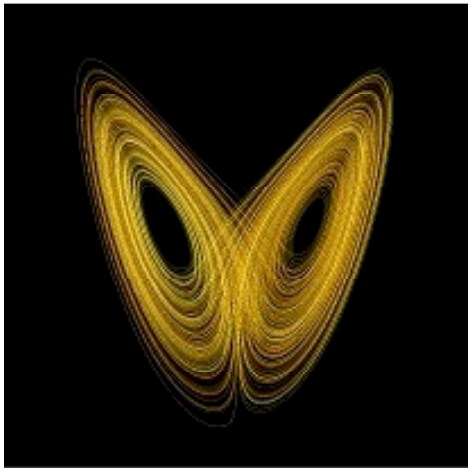
## More complex state spaces — B (two attractors)







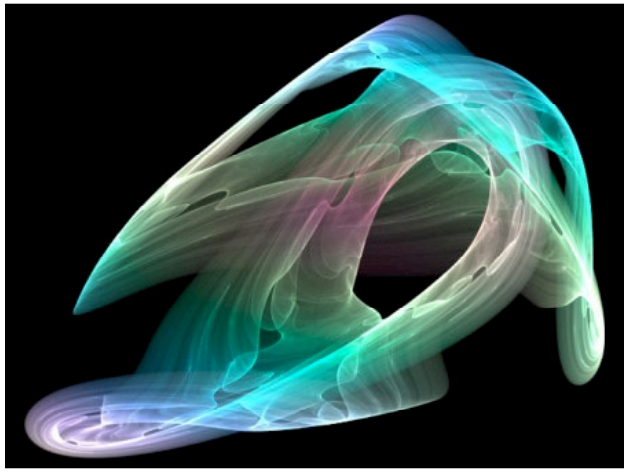

## More complex state spaces — C (two attractors)







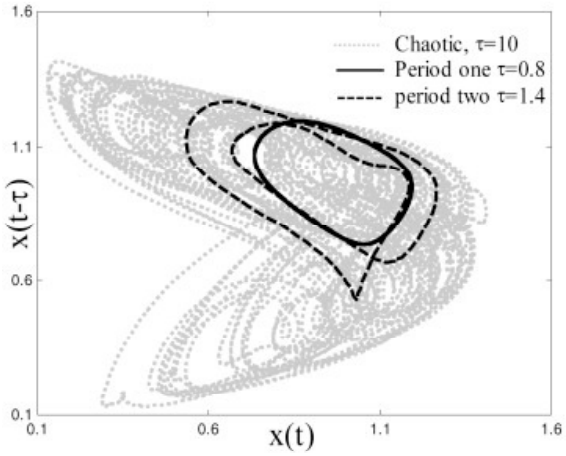
## More complex state spaces — D



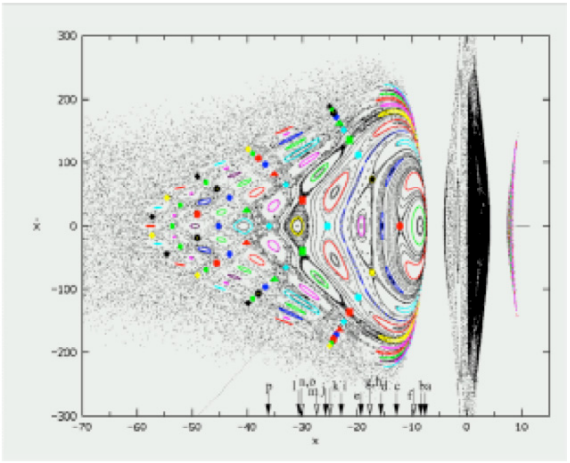




## More complex state spaces — E (chaotic behaviour)



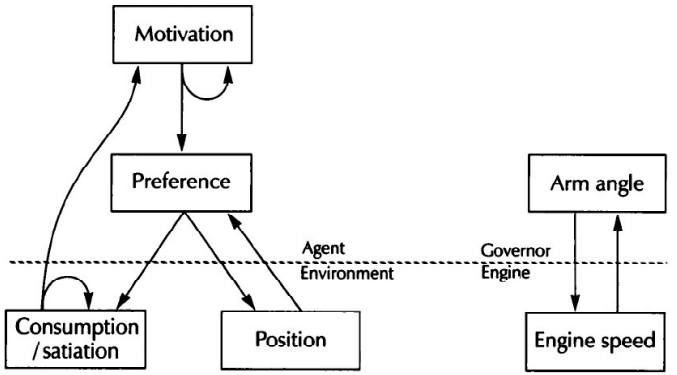
## More complex state spaces — F (chaotic behaviour)







## How does this relate to mind?

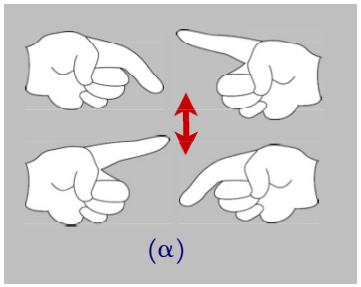


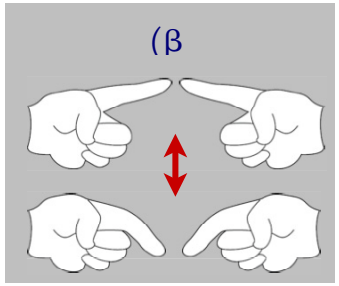





## A famous example

If one starts out wagging one's fingers "out of phase", as in α, and speeds up …




… one inevitably ends up doing it "in phase," as in β





## A famous example (cont'd)

1. This "finger-wagging" behaviour is described by the Haken–Kelso–Bunz (HKB) model of bimanual coordination

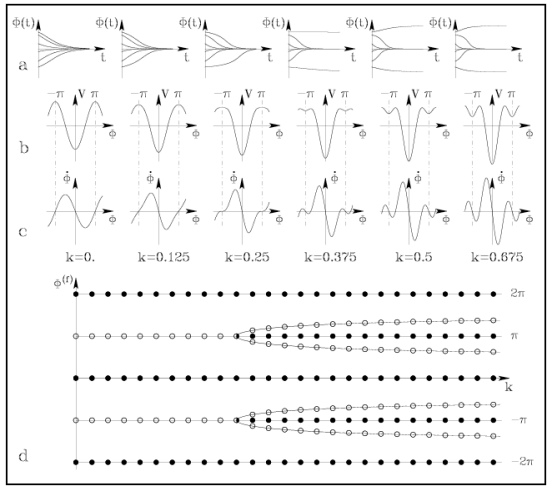
   $V = - a \cos \varphi - b \cos 2\varphi$

   a) φ is the relative phase of the two fingers (0° = in phase; 180° = out of phase)

   NB: φ does not refer to a "part" of any mechanism!

   b) b/a is the frequency

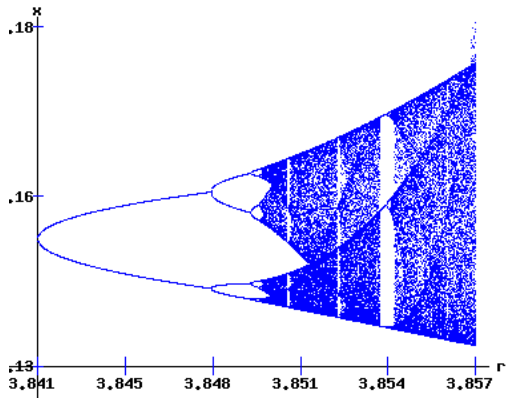
Behaviour of HKB model




Faster ← → Slower












## Van Gelder's comparison of "computational" and dynamical systems

1. (Alleged) properties of the "computational" approach
   a) Representational
   b) "Computational"
   c) Sequential, cyclic operation
   d) Communicating parts

2. (Alleged) properties of the dynamical approach
   a) Non-representational
   b) Non-"computational"
   c) Parallel, continuous operation
   d) No communication (just continuous causal interaction)

Whether GOFAI (let along general computational) systems have "communicating parts" is unclear—in fact it is not entirely clear what this claim even *means*.

By "computational" van Gelder means "as in logic-based GOFAI systems—a much more specific characterization than what real-world computation is actually like.

DST does not theorize systems *as* representational—but it does not *require* that they not be. Ontologically, in fact it is neutral as to whether systems are representational or not. One could as easily use DST to analyze a representational system as a non-representational one—though of course the DST analysis would not deal with its representationality.





## Van Gelder's comparison of "computational" and dynamical systems (cont'd)

1. Dynamical systems are described with **equations**

$$\frac{d^2\theta}{dt^2} = (n\omega)^2 \cos\theta \sin\theta - \frac{g}{l}\sin\theta - r\frac{d\theta}{dt}$$

2. These equations are an (algebraic) part of mathematics known as **dynamical systems theory** (**DST**)
3. Equations include the **environment**
   a) The fact that DST includes the environment is crucial—but so do logic & GOFAI!
   b) DST includes it as a *cause*.
   c) Logic & GOFAI & representational systems include the environment as part of the (normatively governing) *semantic realm*
4. Equations require **numerical properties** ("measure variables")
5. Contrast with logic/GOFAI, which deals in **propositions** and arbitrary-sized **data structures**
6. This distinction between analyzing the parts of a mechanism as **compositional (representational) symbols** and as **items with a scalar (numerical) measure** may ultimately be the most important difference between DST and GOFAI approaches.

## Part III · B

# Dynamical Systems (cont'd)





## van Gelder's 3-way comparison of GOFAI, connectionist, and dynamical systems

Are any of these incompatible with any of the others? *No!*

Or are these? *No!*

| | Computational systems: GOFAI | Connectionist systems | Dynamical systems |
|---|---|---|---|
| Informal description | | | |
| Classic exemplars | | | |
| Kinds of variable | | | |
| Changes in states | | | |
| Tools for description | | | |
| General character | | | |

Table 16.1: Differences among kinds of systems.

Continuous

From Timothy van Gelder, "Dynamics and Cognition", p. 434. in John Haugeland, *Mind Design II*, MIT Press 1996.





## 2-way comparison of GOFAI and dynamical properties

According to van Gelder!

| | | GOFAI Computation | Dynamical Systems | Parallel (coherent) | Opposition? | Best for mind? |
|---|---|---|---|---|---|---|
| 1 | Dynamics | Focus: *states* (Change as motion from one state to another) | Focus: *change* (States of little intrinsic interest, just medium for change) | ✔ | ✘ | ? |
| 2 | Topography | | | ✘ | ✘ | ? |
| 3 | Temporality | | | ✔ | ✘ | ? |
| 4 | Timing vs. order | | | (✔) | ✘ | ? |





## 2-way comparison of GOFAI and dynamical properties (cont'd)

According to van Gelder!

| | | GOFAI Computation | Dynamical Systems | Parallel (coherent) | Opposition? | Best for mind? |
|---|---|---|---|---|---|---|
| 5 | Parallel vs. serial | Serial (most "variables" remain unchanged at each state transition). Change *local*. | Parallel (all aspects change interdependently at the same time). Change *global*. | ✔ | ✔ | ? |
| 6 | Engagement | | | ✔ | (✔) | ? |
| 7 | Interaction | | | ✔ | (?) | ? |
| 8 | Representation | | | (✔) | (✔) | ? |

## Difference in kind of explanation, too

1. **Computation**
   a) Fundamental idea is about **how it works** (mechanism)
   b) Behaviour—i.e, **how it behaves**—is "emergent"
      — Surely behaviour is not in general *surprising*, though (though it may be in particular!) since we typically design them explicitly?
   c) Theories are **mechanical** explanations
2. **Dynamics**
   a) Fundamental idea is about **how it behaves**—i.e., behaviour what it does
   b) The mechanism—i.e., **how it works**—is left *unspecified*
   c) Theories are "Covering law" explanations (vs. mechanism)
      — Cf. Isaac Newton, classical physics "*Hypotheses non fingo*"





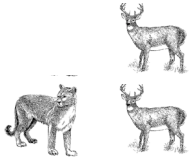
## Other Questions/Issues

1. All DST variables are numeric (measure properties). What are the chances that **mind** will be **explained in numerical terms**?





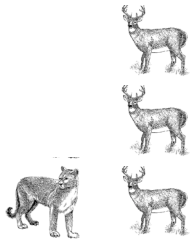
## DST (covering-law) accounts of high-level behaviour/regularities



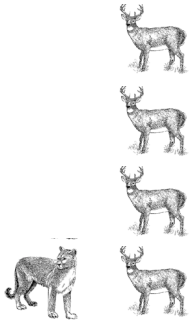




## DST (covering-law) accounts of high-level behaviour/regularities



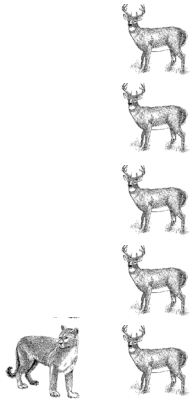
## DST (covering-law) accounts of high-level behaviour/regularities







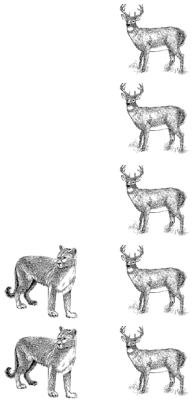
## DST (covering-law) accounts of high-level behaviour/regularities







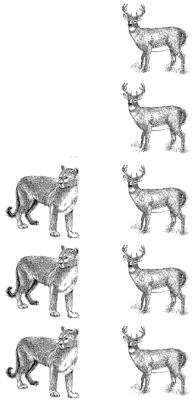
## DST (covering-law) accounts of high-level behaviour/regularities



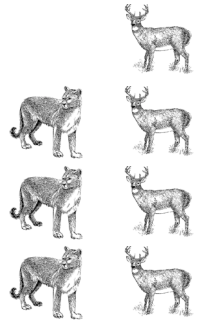




## DST (covering-law) accounts of high-level behaviour/regularities



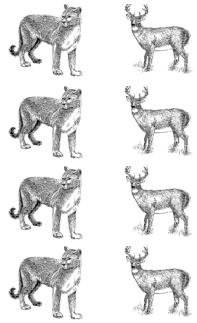
## DST (covering-law) accounts of high-level behaviour/regularities







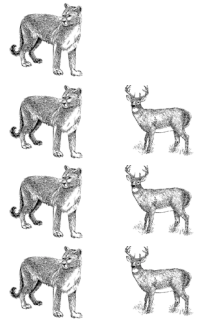
## DST (covering-law) accounts of high-level behaviour/regularities







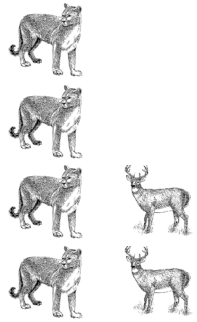
## DST (covering-law) accounts of high-level behaviour/regularities



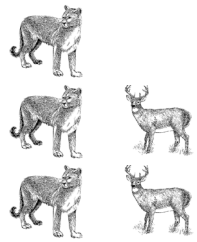




## DST (covering-law) accounts of high-level behaviour/regularities



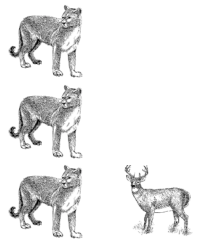
## DST (covering-law) accounts of high-level behaviour/regularities







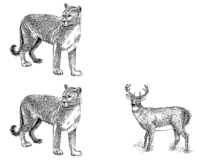
## DST (covering-law) accounts of high-level behaviour/regularities







## DST (covering-law) accounts of high-level behaviour/regularities

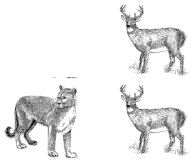






## DST (covering-law) accounts of high-level behaviour/regularities

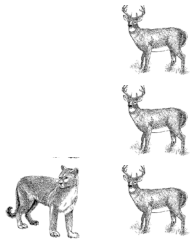
## DST (covering-law) accounts of high-level behaviour/regularities







## DST (covering-law) accounts of high-level behaviour/regularities





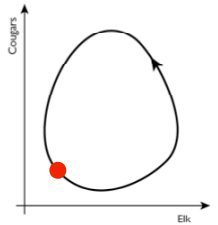


## DST (covering-law) accounts of high-level behaviour/regularities



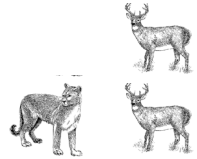


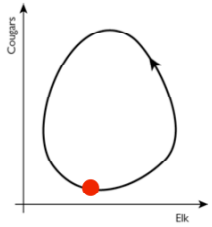
## DST (covering-law) accounts of high-level behaviour/regularities




*State space diagram*

## DST (covering-law) accounts of high-level behaviour/regularities



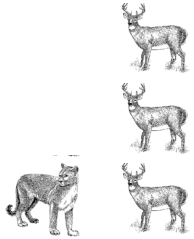


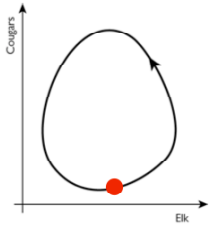

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities



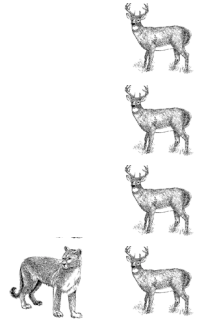


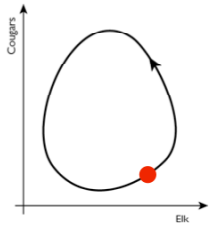

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities



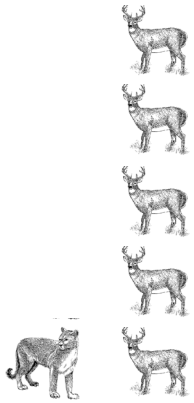


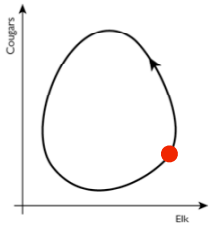

*State space diagram*





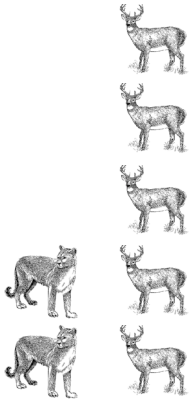
## DST (covering-law) accounts of high-level behaviour/regularities

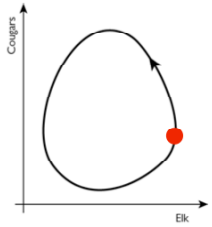





*State space diagram*

## DST (covering-law) accounts of high-level behaviour/regularities



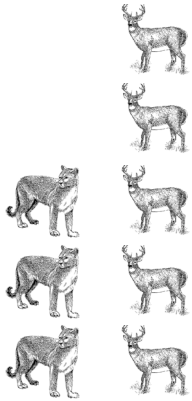


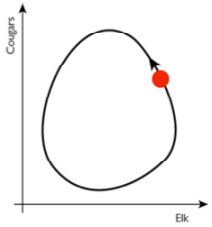

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities



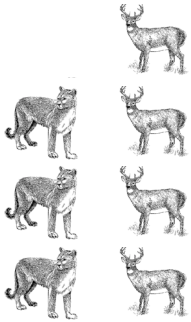


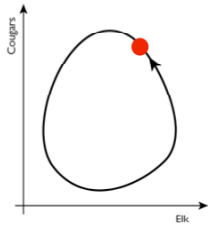

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities



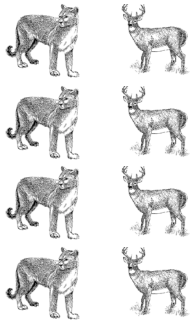


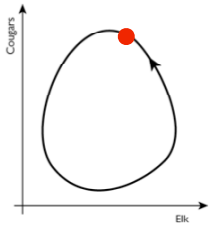

*State space diagram*





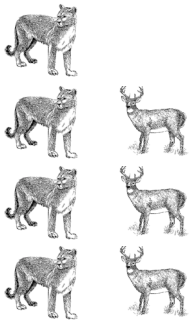
## DST (covering-law) accounts of high-level behaviour/regularities

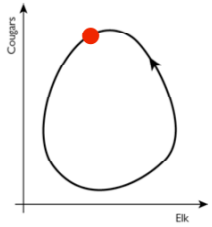





*State space diagram*

## DST (covering-law) accounts of high-level behaviour/regularities



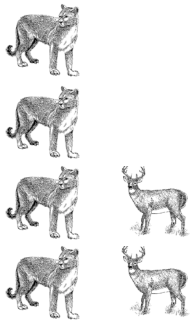


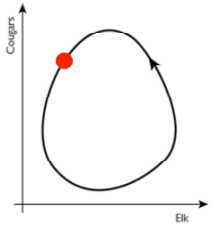

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities



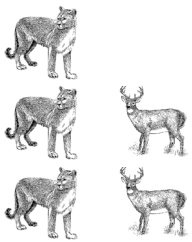


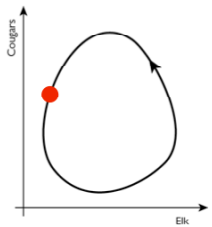

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities



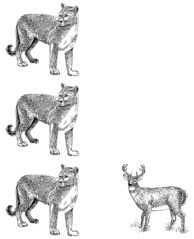


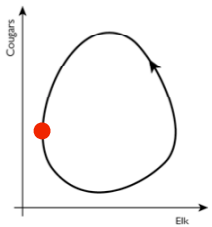

*State space diagram*





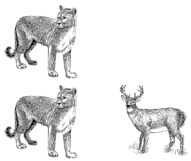
## DST (covering-law) accounts of high-level behaviour/regularities

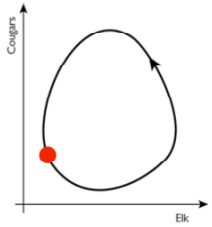





*State space diagram*

## DST (covering-law) accounts of high-level behaviour/regularities



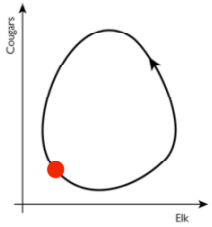



*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities

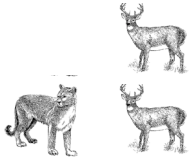


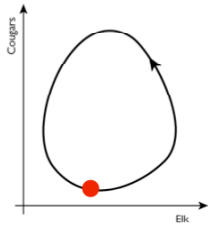

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities



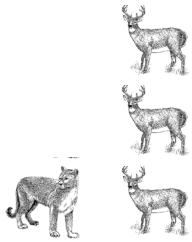


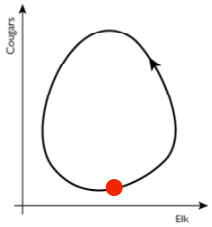

*State space diagram*





## DST (covering-law) accounts of high-level behaviour/regularities






*State space diagram*

## Many other human behaviours are modeled this way

http://demonstrations.wolfram.com/RomeoAndJuliet/
http://sprott.physics.wisc.edu/pubs/paper277.pdf

Sometimes in dubious ways …

https://www.youtube.com/watch?v=SmIxvU3v8t4





## In Sum

1. As with connectionism / neural networks, an exploration of dynamical systems shows that we are just beginning to explore the contours of the whole design space.
   a) Not just the design space of **how these systems work**
   b) But also the space of **possible ways of explaining them** (more on this in a moment).
2. In addition, if one realizes that computation-in-the-wild is not the same thing as GOFAI, these considerations lead one to wonder whether this evolution of theoretical ideas (GOFAI, dynamic systems, etc.) is not more

✔ a) **Epistemological** (how we understand systems) than
   b) **Ontological** (what the system is, how the system works)





## DST as a different way of looking at systems

— *I.e., as more of an epistemological than ontological difference*

| | | GOFAI Computation | Dynamical Systems | Parallel (coherent) | Opposition? | Best for mind? |
|---|---|---|---|---|---|---|
| 1 | Dynamics | Focus: *states* (Change as motion from one state to another) | Focus: *change* (States of little intrinsic interest, just medium for change) | ✔ | ✘ | ? |
| 2 | Topography | *Internal combinatoric syntactic structure* (how smaller pieces contribute to larger wholes) | Understand system *geometrically*, in terms of positions in state space (dynamical landscape) | ✘ | ✘ | ? |
| 3 | Temporality | Laid out *statically*—all present at one time. Cognition as transformations of static structures | Laid out *dynamically*: simultaneous, mutual unfolding of complex temporal patterns | ✔ | ✘ | ? |
| 4 | Timing vs. order | What behaviour is, regardless of timing details. *Which* states a system passes through | How behaviours happen in time (?) *When* system passes through states | (✔) | ✘ | ? |

— Yellow = *epistemological* (a way of *looking at* a system)





## DST as a different way of looking at systems

— *I.e., as more of an epistemological than ontological difference*

| | | GOFAI Computation | Dynamical Systems | Parallel (coherent) | Opposition? | Best for mind? |
|---|---|---|---|---|---|---|
| 5 | Parallel vs. serial | Serial (most "variables" remain unchanged at each state transition). Change *local*. | Parallel (all aspects change interdependently at the same time). Change *global*. | ✔ | ✔ | ? |
| 6 | Engagement | Start with *input*, produce appropriate *output* (via sequence of internal operations; halt on output). | Processes ongoing—not starting anywhere, not finishing anywhere. Goal not to map input onto output, but to constantly maintain appropriate change. | ✔ | (✔) | ? |
| 7 | Interaction | *Sees state.* System changes in its own way from that state, until new input resets state again | Input as ongoing parameter influence on *shape of change*. Output as ongoing influence on something else. Sometimes *coupling*—two systems simultaneously shaping each other's change. | ✔ | (?) | ? |
| 8 | Representation | Computation inherently representational. Cognition depends on manipulations of internal representation. Representations as static configurations of symbol tokens. | Dynamical systems *not* inherently representational. Some: representations in parameter settings, system states, attractors, trajectories, or even aspects of bifurcation structures. Some (small but influential): notion of representation *dispensable* or *hindrance* | (✔) | (✔) | ? |

## In Sum

1. As with connectionism / neural networks, an exploration of dynamical systems shows that we are just beginning to explore the contours of the whole design space.
   a) Not just the design space of **how these systems work**
   b) But also the space of **possible ways of explaining them** (more on this in a moment).
2. In addition, if one realizes that computation-in-the-wild is not the same thing as GOFAI, these considerations lead one to wonder whether this evolution of theoretical ideas (GOFAI, dynamic systems, etc.) is not more
   ✔ a) **Epistemological** (how we understand systems) than
   b) **Ontological** (what the system is, how the system works)
3. Dynamical systems theory (DST) explanations put an emphasis on:
   a) Dynamic temporal behaviour (how things **behave** and **change**) of
   b) Parallel, continuous behavioural components
   c) Numerically measured (expressed in **differential equations**)
   d) Analyzed in terms of **state spaces** of possible behavioural configurations
   e) Including the state of the **environment**





## In Sum (cont'd)

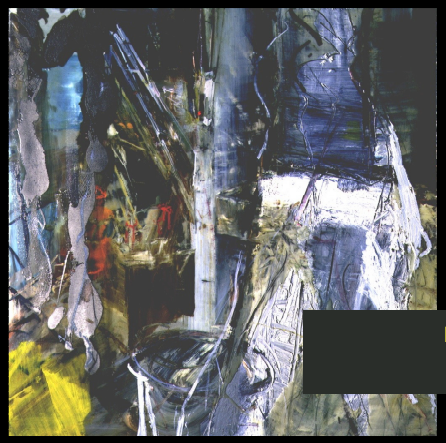
4. DST does *not* put any emphasis on:
   a) Underlying mechanism (how things **work**)
   b) Compositional **states**
   c) **Representational** capacities
5. Thinking of symbol manipulation or networks or dynamical systems as *ideologically-opposed alternatives* (the way van Gelder presents them!) is probably not the most productive approach.
6. Instead: We should take seriously all these issues: dynamics, learning, networks, continuity, representation, etc., and explore their myriad forms of combination…

**Same conclusion as for neural networks:**





*In place of the regular format, this lecture is (a version of) a talk that I gave on at the UofT Ethics Centre on how to understand what sorts of intelligence or "cognitive capacity" we can expecct from deep learning systems, and what kinds we cannot.*



Reckoning and Judgment

Brian Cantwell Smith



## Part I — History

The four (vaguely Cartesian) assumptions of **GOFAI**

"Good old-fashioned Artificial Intelligence" — Haugeland

1. The essence of intelligence is **thought**
2. The epitome of thought is **logical inference**
3. **Perception**, at a lower level than thought, won't be that hard
4. "**Formal ontology**": world consists of discrete objects and "clear and distinct" properties—evident in the vocabulary of natural language





## GOFAI was also based on a broader general insight:

It is possible to construct a physical system that:

1. **Works**, mechanically, on straightforward physical principles (amenable to science)
2. Is **semantically interpretable**—has behaviour intelligible in terms of relations of meaning and reference to the external world
   a. Implying a distinction between *what they do* and *how they work*
3. The semantic reference relations are **not effective** (not causal, making them impossible to "detect", and implying that they are not explicable in science).
4. **Normatively governed** in terms of its semantic interpretation





## GOFAI was also based on a broader general insight:

It is possible to construct a physical system that:

1. **Works**, mechanically, on straightforward physical principles (amenable to science)
2. Is **semantically interpretable**—has behaviour intelligible in terms of relations of meaning and reference to the external world
   a. Implying a distinction between *what they do* and *how they work*
3. The semantic reference relations are **not effective** (not causal, making them impossible to "detect", and implying that they are not explicable in science).
4. **Normatively governed** in terms of its semantic interpretation

## GOFAI was also based on a broader general insight:

It is possible to construct a physical system that:

1. **Works**, mechanically, on straightforward physical principles (amenable to science)
2. Is **semantically interpretable**—has behaviour intelligible in terms of relations of meaning and reference to the external world
   a. Implying a distinction between *what they do* and *how they work*
3. The semantic reference relations are **not effective** (not causal, making them impossible to "detect", and implying that they are not explicable in science).
4. **Normatively governed** in terms of its semantic interpretation





## GOFAI was also based on a broader general insight:

It is possible to construct a physical system that:

1. **Works**, mechanically, on straightforward physical principles (amenable to science)
2. Is **semantically interpretable**—has behaviour intelligible in terms of relations of meaning and reference to the external world
   a. Implying a distinction between *what they do* and *how they work*
3. The semantic reference relations are **not effective** (not causal, making them impossible to "detect", and implying that they are not explicable in science).
4. **Normatively governed** in terms of its semantic interpretation





## GOFAI was also based on a broader general insight:

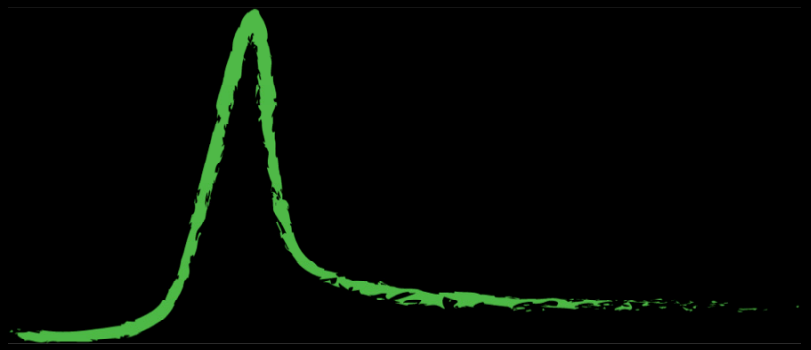
It is possible to construct a physical system that:

1. **Works**, mechanically, on straightforward physical principles (amenable to science)
2. Is **semantically interpretable**—has behaviour intelligible in terms of relations of meaning and reference to the external world
   a. Implying a distinction between *what they do* and *how they work*
3. The semantic reference relations are **not effective** (not causal, making them impossible to "detect", and implying that they are not explicable in science).
4. **Normatively governed** in terms of its semantic interpretation

*Semantics is* **deferential**







*Every great idea languishes for most of history in obscurity, has one brief moment of glory, and then lives out its dying days as a platitude…*

## Nevertheless, GOFAI failed … *or anyway is deemed to have failed*

The four (vaguely Cartesian) assumptions of GOFAI

1. The essence of intelligence is **thought**
2. The epitome of thought is **inference**
3. **Perception**, at a lower level than thought, won't be that hard
4. "**Formal ontology**": world consists of discrete objects and "clear and distinct" properties—evident in the vocabulary of natural language





## Primary critiques of GOFAI

1. **Psychological:** the brain doesn't work that way
2. **Perceptual:** the world is a *mess!*
3. **Ontological:** the world doesn't come neatly chopped up into objects
4. **Epistemological**
   a. Intelligence doesn't (in general) consists of rational, articulated, steps; it is better understood as patterns of skilful navigation and coping—being "thrown" into enmeshing social and personal projects
   b. Thinking emerges from an unconscious background—a horizon of ineffable knowledge and sense-making
   c. **Commonsense!** ("take out the kidney and boil it")





## Primary critiques of GOFAI

1. **Psychological:** the brain doesn't work that way
2. **Perceptual:** the world is a *mess!*
3. **Ontological:** the world doesn't come neatly chopped up into objects
4. **Epistemological**
   a. Intelligence doesn't (in general) consists of rational, articulated, steps; it is better understood as patterns of skilful navigation and coping—being "thrown" into enmeshing social and personal projects
   b. Thinking emerges from an unconscious background—a horizon of ineffable knowledge and sense-making
   c. **Commonsense!** ("take out the kidney and boil it")



You just processed this image using a neuronal device comprising 100 billion elements with 100 trillion interconnections honed for this explicit purpose over 500 million years of evolution!



## Primary critiques of GOFAI

1. **Psychological:** the brain doesn't work that way
2. **Perceptual:** the world is a *mess!*
3. **Ontological:** the world doesn't come neatly chopped up into objects
4. **Epistemological**
   a. Intelligence doesn't (in general) consists of rational, articulated, steps; it is better understood as patterns of skilful navigation and coping—being "thrown" into enmeshing social and personal projects
   b. Thinking emerges from an unconscious background—a horizon of ineffable knowledge and sense-making
   c. **Commonsense!** ("take out the kidney and boil it")





## Primary critiques of GOFAI

1. **Psychological:** the brain doesn't work that way
2. **Perceptual:** the world is a *mess!*
3. **Ontological:** the world doesn't come neatly chopped up into objects
4. **Epistemological**
   a. Intelligence doesn't (in general) consists of rational, articulated, steps; it is better understood as patterns of skilful navigation and coping—being "thrown" into enmeshing social and personal projects
   b. Thinking emerges from an unconscious background—a horizon of ineffable knowledge and sense-making
   c. **Commonsense!** ("take out the kidney and boil it")





## Primary critiques of GOFAI

1. **Psychological:** the brain doesn't work that way
2. **Perceptual:** the world is a *mess!*
3. **Ontological:** the world doesn't come neatly chopped up into objects
4. **Epistemological**
   a. Intelligence doesn't (in general) consists of rational, articulated, steps; it is better understood as patterns of skilful navigation and coping—being "thrown" into enmeshing social and personal projects
   b. Thinking emerges from an unconscious background—a horizon of ineffable knowledge and sense-making
   c. **Commonsense!** ("take out the kidney and boil it")

A metaphor about thought

Conceptual content …



## Primary critiques of GOFAI

1. **Psychological:** the brain doesn't work that way
2. **Perceptual:** the world is a *mess!*
3. **Ontological:** the world doesn't come neatly chopped up into objects
4. **Epistemological**
   a. Intelligence doesn't (in general) consists of rational, articulated, steps; it is better understood as patterns of skilful navigation and coping—being "thrown" into enmeshing social and personal projects
   b. Thinking emerges from an unconscious background—a horizon of ineffable knowledge and sense-making
   c. **Commonsense!** ("take out the kidney and boil it")



A metaphor about thought

Beautiful, but not "clear and distinct"

… undergirded by nonconceptual connections

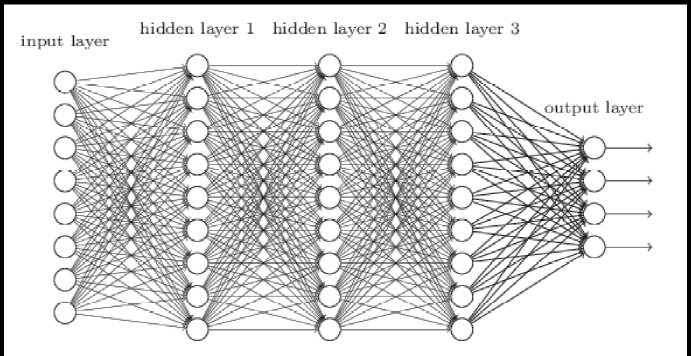


## Primary critiques of GOFAI

1. **Psychological:** the brain doesn't work that way
2. **Perceptual:** the world is a *mess!*
3. **Ontological:** the world doesn't come neatly chopped up into objects
4. **Epistemological**
   a. Intelligence doesn't (in general) consists of rational, articulated, steps; it is better understood as patterns of skilful navigation and coping—being "thrown" into enmeshing social and personal projects
   b. Thinking emerges from an unconscious background—a horizon of ineffable knowledge and sense-making
   c. **Commonsense!** ("take out the kidney and boil it")

## Part II — Deep Learning

1. What is it?
2. What is it capable of?
3. What should we make of it?








## Four facts about Deep Learning

1. Neural Networks — involve
   a) **Shallow** (few step) inference
   b) On **massive amounts** of data
   c) Involving **very large numbers** of
   d) **Weakly correlated** variables

   Logic (GOFAI) systems — involve
   a) **Deep** (many step) inference
   b) On **modest amounts** of information
   c) Involving a **small number** of
   d) **Strongly correlated** variables
2. Works at an (ineffable) level sorting and sifting **massive amounts of data**
3. Can **learn**—be **trained** ⟸ a holy grail of AI
4. In training, uses a **phenomenal amount of computational power**





## Four facts about Deep Learning

1. Neural Networks — involve
   a) **Shallow** (few step) inference
   b) On **massive amounts** of data
   c) Involving **very large numbers** of
   d) **Weakly correlated** variables

   Logic (GOFAI) systems — involve
   a) **Deep** (many step) inference
   b) On **modest amounts** of information
   c) Involving a **small number** of
   d) **Strongly correlated** variables
2. Works at an (ineffable) level sorting and sifting **massive amounts of data**
3. Can **learn**—be **trained** ⟸ a holy grail of AI
4. In training, uses a **phenomenal amount of computational power**





## Four facts about Deep Learning

1. Neural Networks — involve
   a) **Shallow** (few step) inference
   b) On **massive amounts** of data
   c) Involving **very large numbers** of
   d) **Weakly correlated** variables

   Logic (GOFAI) systems — involve
   a) **Deep** (many step) inference
   b) On **modest amounts** of information
   c) Involving a **small number** of
   d) **Strongly correlated** variables
2. Works at an (ineffable) level sorting and sifting **massive amounts of data**
3. Can **learn**—be **trained** ⟸ a holy grail of AI
4. In training, uses a **phenomenal amount of computational power**

## Four facts about Deep Learning

| Neural Networks — involve | Logic (GOFAI) systems — involve |
|---|---|
| a) **Shallow** (few step) inference | a) **Deep** (many step) inference |
| b) On **massive amounts** of data | b) On **modest amounts** of information |
| c) Involving **very large numbers** of | c) Involving a **small number** of |
| d) **Weakly correlated** variables | d) **Strongly correlated** variables |

2. Works at an (ineffable) level sorting and sifting **massive amounts of data**
3. Can **learn**—be **trained** ⟸ a holy grail of AI
4. In training, uses a **phenomenal amount of computational power**

The results are extraordinarily impressive!





## Which of the GOFAI critiques does Deep Learning deal with?

| | | |
|---|---|---|
| ✔ | 1. **Psychological** | ≈ pretty well |
| | 2. **Perceptual** | |
| | 3. **Ontological** | |
| | 4. **Epistemological** | |





## Which of the GOFAI critiques does Deep Learning deal with?

| | | |
|---|---|---|
| ✔ | 1. **Psychological** | ≈ pretty well |
| ✔ | 2. **Perceptual** | ≈ pretty well |
| | 3. **Ontological** | |
| | 4. **Epistemological** | |





## Which of the GOFAI critiques does Deep Learning deal with?

| | | |
|---|---|---|
| ✔ | 1. **Psychological** | ≈ pretty well |
| ✔ | 2. **Perceptual** | ≈ pretty well |
| ✔ | 3. **Ontological** | ≈ yes |
| | 4. **Epistemological** | |

## Which of the GOFAI critiques does Deep Learning deal with?

- ✔ 1. **Psychological** ≈ pretty well
- ✔ 2. **Perceptual** ≈ pretty well
- ✔ 3. **Ontological** ≈ yes
- 4. **Epistemological** *... not so fast!*





## Limits of Deep Learning

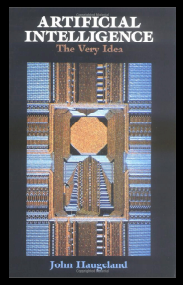
1. In a very important sense, the DL systems (at least any that have been described so far) **don't know what they are talking about**.
2. They don't know the difference (can't "tell" the difference) between
   a. Their own states (and the states of their inputs and outputs)
   b. The state of the world their states (and inputs & outputs) represent
3. What founds semantics—what founds *understanding*—is **deferential commitment** to the world—the world we are *in*, *of*, and *about*.
4. Though we may *design* them with deferential semantics, in all systems that have been described, **the deference is ours, not theirs**
5. Therefore the objects in the world that they deal with aren't really *objects* for them—or even *in the world*, for them.
6. To build systems that truly understand, that are genuinely intelligent, we have to construct systems that are **themselves deferential**—that themselves *submit* to the worlds they inhabit.

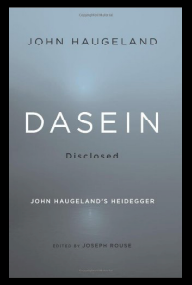


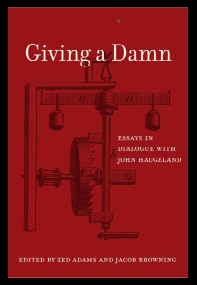


## Part III — Deference

*I owe much of my thinking on these issues to John Haugeland (who passed away before I could talk to him about framing the AI debate in terms of* deference *and* judgment*)*








## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**

## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**





## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**





## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**





## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**

## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**





## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**





## Part III — Deference

### Conditions on taking an object to be an object

1. **Orientation/comportment**
2. Distinction between **appearance and reality**
3. **Intelligibility** (in terms of the rules & regularities of a constituting regime)
4. Difference between **right and wrong**
   a. What is the case (right)
   b. What is not the case (wrong)
   c. What couldn't be the case (impossible)
5. Existential **commitment**
6. Epistemic **self-awareness**
7. The **world**





### Animals, reckoning, and judgment

## Part IV — Ethics

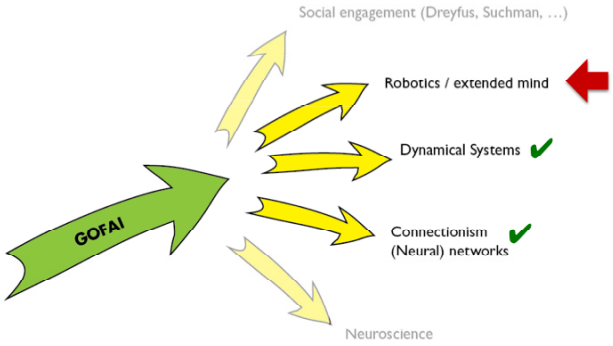
The biggest threat that AI poses to humanity and the world is that we will hand over, to systems that merely **reckon**, responsibility for matters whose stewardship demands *passionate, dispassionate, compassionate* **judgment**.





## Part III · C
# Embodied Robotics

Nov 14, 2017








## GOFAI robotics (1960s and 1970s)

1. Used plans, goals, high-level representations, central-level control
2. Planning and perception programs ran on a mainframe (typically a DEC PDP-10); motor commands broadcast to robot, and sensors signals broadcast back
3. PDP-10 statistics
   a) Machine cost: ~$1,000,000
   b) Memory: 300,000 (300K) 36-bit words
      i. Cost in 1969: **$1,000,000.00** (1969 dollars)
      ii. Cost in 1969: **$7,000,000.00** (2017 dollars)
      iii. Cost in 2017: **$0.005 (½ cent)** (2017 dollars)
      — price reduction of ~99.9999999% (i.e., drop in price of ~$10^9$!)
   c) Processing power: **~500 KLOPS** (~one millionth of the power of an iPhone X!)





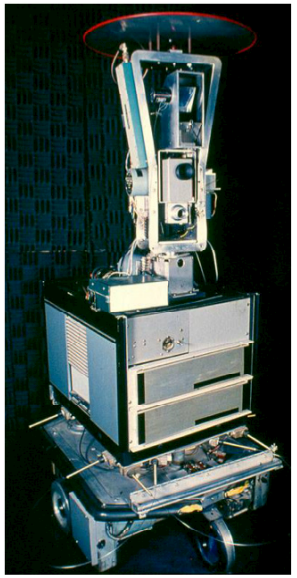
## Digital Equipment Corporation (DEC) PDP-10





## Shakey @ SRI — 1969

https://www.youtube.com/watch?v=GmU7SimFkpU



## Rod Brooks: "Intelligence without Representation" (1987/1991)

1. Start over! The GOFAI approach is *completely backwards*!
2. Don't try to build what we humans think an "intelligent robot" would be like.
3. Instead, incrementally build up the capabilities of intelligent systems, constructing *complete systems* at each step of the way and thus automatically ensure that the pieces and their interfaces are valid.
4. Let each stage of complete intelligent systems loose in the real world, with real sensing and real action. Anything less provides a candidate *with which we can delude ourselves*.





## Rod Brooks: "Intelligence without Representation" (1987/1991)

1. Start over! The GOFAI approach is *completely backwards*!
2. Don't try to build what we humans think an "intelli… be like.
3. Instead, incrementally build up the capabiliti… intelligent systems, constructing *complete…* at each step of the way and thus au… ensure that the pieces and thei… d.
4. Let each stage of compl… ms loose in the real world, w… d real action. Anything less… te *with which we can delude ou…*

"Today the earwig; tomorrow, man"

David Kirsh, *Artificial Intelligence* 47 (1991) 161-184





## Observation and Hypothesis

**O** • When we examine very simple level intelligence *we find that explicit representations and models of the world simply get in the way*. It turns out to be better to **use the world as its own model.**

**H** • Representation is the wrong unit of abstraction in building **the bulkiest parts** of intelligent systems.

Brooks' most famous line!

Very interesting; Very telling





## Timing (years)

| | |
|---|---|
| ~3,500,000,000 | single cell organisms |
| ~2,500,000,000 | photo-synthetic plants |
| ~550,000,000 | first fish and vertebrates |
| ~330,000,000 | dinosaurs |
| ~250,000,000 | mammals |
| ~125,000,000 | primates |
| ~2,500,000 | homo sapiens |
| ~10,000 | agriculture |
| ~5,000 | writing |
| ~300 | "expert" knowledge |

Concentrate here! (says Brooks) This is >99.9% of our evolution

## An airplane metaphor ...





## Brooks: quotes

1. "There is no clean division between **perception** (abstraction) and **reasoning** in the real world." ✔
2. "**Abstraction** is the essence of intelligence and the hard part of the problems being solved."

*A similar sentiment to the views of supporters of machine learning and neural-network architectures*





## Criteria on "a Creature" — Negative ✘

1. No interest in **how people work** ( *not so good for us cognitive scientists*)
2. No interest in **applications** (initially!) ( *that's OK* )
3. No interest in **philosophical implications** ( *we'll fix that!* )





## Criteria on "a Creature" — Positive ✔

1. Must **cope appropriately and in a timely fashion** with changes in its dynamic environment.
2. Should be **robust with respect to its environment**; minor changes in the properties of the world should not lead to total collapse of the Creature's behaviour; rather one should expect only a gradual change in capabilities of the Creature as the environment changes more and more.
3. Should be able to **maintain multiple goals** and, depending on the circumstances it finds itself in, change which particular goals it is actively pursuing; thus it can both adapt to surroundings and capitalize on fortuitous circumstances.
4. Should **do something in the world**; it should have some **purpose** in being.

## Subsumption architecture — "behaviour-based systems"

1. Decompose
   a) By **activity** (e.g., "avoiding walls," "escape danger")
   b) Rather than by **function** (e.g., perception/action/reasoning)
2. *No (central?) processing* (contrast Shakey)
3. *No central representation*
4. **Layers of parallel activity**, connected by **suppression** and **inhibition**

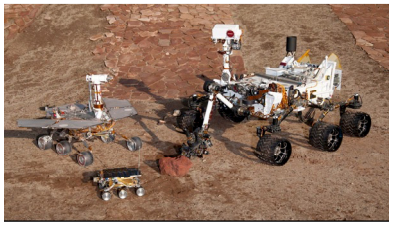




## Videos

✔ 1. **Brooks — Robots** · http://www.youtube.com/watch?v=C9p8B7-5MTI
✔ 2. **Brooks — 2003 (TED)** · http://www.youtube.com/watch?v=UdyRmdv-KiY
3. **Brooks — 2008** · http://www.youtube.com/watch?v=O5DIyUWR-YY

3 generations of Mars Rovers (on which Brooks consulted)







## Important properties

| | | |
|---|---|---|
| **1. Situatedness** | World is its own best model | (✔) *Yes; but it is not always available* |
| **2. Embodiment** | World grounds the regress of meaning-giving | ✔ Yes, for sure |
| **3. Intelligence** | Determined by *dynamics* of *behaviour* (behaviourist?) | ~ Overall probably so; locally, no |
| **4. Emergence** | Intelligence is in the eye of the beholder | ✘ No (but why does he need this?) |





## Examples of follow-on work (much by Marc Raibert)

1. **Boston Dynamics — Rise**
   · http://www.youtube.com/watch?v=YPQ25TOHTXk
2. **Locust-inspired jumping robot**
   · http://www.youtube.com/watch?v=ADiHexd3UcY
3. **MIT hopping robot**
   · http://www.youtube.com/watch?v=XFXj81mvInc
4. **Boston Dynamics — Big Dog (demo)**
   · http://www.youtube.com/watch?v=cNZPRsrwumQ
5. **Boston Dynamics — Big Dog (explanation)**
   · http://www.youtube.com/watch?v=-Bi-tPO0OPs
6. **Boston Dynamics — Spot (vs. Fido)**
   · https://www.youtube.com/watch?v=S7nhygaGOmo
7. **Boston Dynamics — Petman**
   · http://www.youtube.com/watch?v=4eJAm_MY698&
8. **Boston Dynamics — Resilience**
   · https://www.youtube.com/watch?v=4PaTWufUqqU

## Cheetah

1. **The Robot**
   - http://www.youtube.com/watch?v=chPanW0QWhA
2. **The Gold Standard**
   - http://vimeo.com/53914149





## Critiques

1. **Cheating**
   a) *Challenge*: Ingenuity is in Brooks, Raibert, students — not robot
   b) *Brooks' reply*: Lots of ingenuity in evolution, too
   c) *Issue*: Suppose ingenuity made us. What should cogsci study:
      i) The ingenuity that made us? *or*
      ii) What that ingenuity made us into?
2. **Scale**
   a) *Challenge*: Will it scale up?
   b) *Brooks' reply*: Send money!
   c) *Issue*: Don't all the classical critiques of behaviourism apply? (This is tricky, because they *do* see inside; but they are *evaluating* and *judging* it from the outside, purely behaviourally. This is especially evident in Kermit and the "emotional" robots.)





## Critiques (cont'd)

3. **Implications**
   a) *Challenge:* So what? What about cognition?
   b) *Brooks' reply:* i) Our robots do 95% of what humans do;
      ii) Perception is most of intelligence;
      iii) Only if we do everything relying on perception first will we know what we need in addition
   c) *Issue:* Cf. below…
4. **Representation**
   a) *Challenge:* Aren't you going to need representation at some point?
   b) *Brooks' reply:* There is representation in the layers (distributed)
   c) *Issue:* … **we need to talk** …

Brooks' changed his tune!





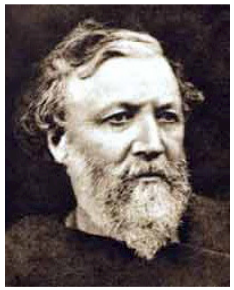
## Representational challenges

1. Some things it seems that a Brooksian robot could not do:
   a) *Entertain hypotheticals*
   b) *Report on what it has done*
   c) *Plan sensibly*
   d) *Make deliberate adjustments to its behaviour*
   e) *Receive instructions and considerations from others*
2. "Perceptible world" is not a good model of
   a) What the world would be like, *if things were different*
   b) How things were, *a moment ago*
   c) How things are going to be, *in the future*
   d) What the world looks like *from a different perspective*
3. What is "directly perceptible" is *that to which you are connected*
4. What representation is good for is to let you know how things are *to which you are not currently connected*

We'll talk about connection & disconnection in the next lecture

## Part III · C
## Embodied Robotics (cont'd)

**R. Brooks**
**v.**
**R. Browning**



Nov 16, 2017





## Critiques

1. **Cheating**
   - a) *Challenge*: Ingenuity is in Brooks, Raibert, students — not robot
   - b) *Brooks' reply*: Lots of ingenuity in evolution, too
   - c) *Issue*: Suppose ingenuity made us. What should cogsci study:
     - i) The ingenuity that made us? *or*
     - ii) What that ingenuity made us into?
2. **Scale**
   - a) *Challenge*: Will it scale up?
   - b) *Brooks' reply*: Send money!
   - c) *Issue*: Don't all the classical critiques of behaviourism apply? (This is tricky, because they *do* see inside; but they are *evaluating* and *judging* it from the outside, purely behaviourally. This is especially evident in Kermit and the "emotional" robots.)





## Critiques (cont'd)

3. **Implications**
   - a) *Challenge:* So what? What about cognition?
   - b) *Brooks' reply:*
     - i) Our robots do 95% of what humans do;
     - ii) Perception is most of intelligence;
     - iii) Only if we do everything relying on perception first will we know what we need in addition
   - c) *Issue:* Cf. below...
4. **Representation**
   - a) *Challenge:* Aren't you going to need representation at some point?
   - b) *Brooks' reply:* There is representation in the layers (distributed)
   - c) *Issue:* ... we need to talk ...

Brooks' changed his tune!





## Clarification — Brooksian Robotics vs. GOFAI

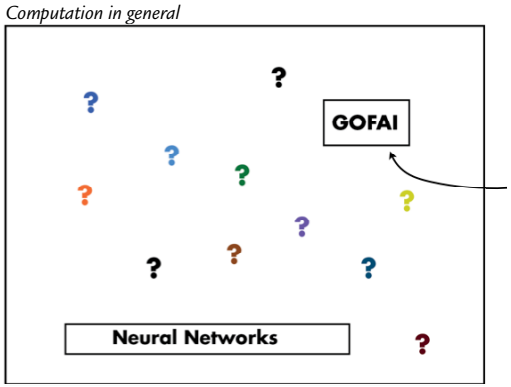
A. Question
   - a) Brooks' robots work by *running computer programs on conventional computer hardware* (as well as using fancy sensors and effectors)
   - b) Does that mean that his robots are really GOFAI systems after all?
   - c) **No!**

B. As we've seen
   - a) In philosophy of mind (and parts of cognitive science) it is common to equate:
     - i. The "**computational theory of mind**", and
     - ii. **GOFAI** ("good old fashioned symbol systems")
   - b) That is, in these fields it is common to assume that *all computer systems involve the formal manipulation of semantically interpreted symbols*, as in logic
   - c) For example: this is what Dreyfus assumes, in mounting his critiques

C. In fact, however
   - a) The space of possible computational architectures† is *vastly larger* than the sorts of "formal symbol manipulation" imagined in this equivalence claim
   - b) Hence the diagram on slides 6 of lecture 08 (a).

*† Even those running on conventional computer architectures!*

## GOFAI as just *one possible kind* of computational architecture†




† *From slide 6 of lecture 08 (a)*





## Clarification — Brooksian Robotics vs. GOFAI (cont'd)

D. A GOFAI system is one that
   a) Treats the activity of the system as "reasoning" about an external world or task domain ⟸ *On the model of logical inference*
   b) Where the reasoning process is implemented by two components:
      i. An inference system, operating over
      ii. A "model" of the task domain
   c) Where the model is represented in a symbol system, called a "representation language", that
   d) Is **systematic**, **productive**, and **compositional**†

E. The programs that power Brooks' robots have *none of these properties*
   a) What the programs do is to *cause physical behaviour*, not *reason*
   b) The behaviour does not result from *inference* or *reasoning* ⟸ *It doesn't take "true sentences" as input, and produce "true sentences" as output*
   c) The programs aren't built to use *models of their task domains*
   d) Though they use data structures, there *isn't a "representation language" with properties of systematicity, productivity, and compositionality.*

† *As defined on slides 19–21 of Lecture 02 (a).*





## Clarification — Brooksian Robotics vs. GOFAI (cont'd)

F. Moral
   a) Be careful when you hear the phrase: the "**computational theory of mind**"
   b) It can mean either of two things
      α A GOFAI system—from people who (mistakenly) think that all computer systems are GOFAI systems (likely philosophers, and non-computational cognitive scientists); or
      β What a programmer would call a "computational system"
   c) As we've seen, category β is much, much bigger than category α





## Challenges to the Brooksian Architecture

1. Some things it seems that a Brooksian robot could **not** do:
   a) Entertain hypotheticals ⟸ *no fantasy lives for Brooksian robots!*
   b) Report on what it has done ⟸ *might be good as criminals*
   c) Plan sensibly *… but dubious as accomplices*
   d) Make deliberate adjustments to its behaviour ⟸ *unlikely to be embarrassed, or to stop doing embarrassing things*
   e) Receive (verbal?) instructions and considerations from others ⟸ *"Never treat your creator like that again!" – won't help*

## Challenges (cont'd)

2. To do any of these things, it is clear that

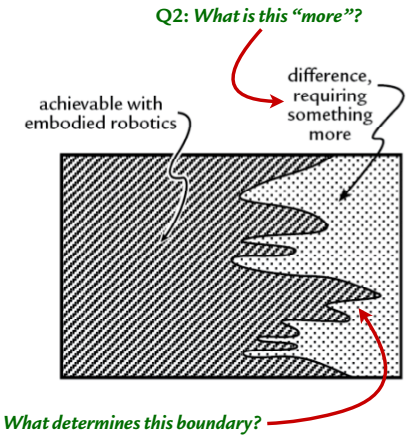
   **Something more is needed …**

   a) What is the "something more"? To answer that requires that we understand:

      i. Why these robots **can do what they can do**

      ii. Why they **can not do what they cannot do**

   b) Leads to two questions








## Challenges (cont'd)

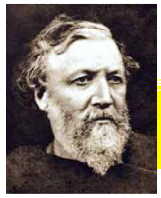
3. The "perceptible world" is not a good model of
   a) **Remote:** how things are *a long way away* (other side of moon)
   b) **Past:** how things were, *a moment ago*
   c) **Future:** how things are going to be, *in the future*
   d) **Alterity:** what the world looks like *from a different perspective*
   e) **Non-existence:** what world would be like, *if things were different*
   f) **Non-effective:** properties (even **local**) that *can't be causally detected*
      i. O'clock properties ("being 2:23 p.m.", "being Nov 17, 2015")
      ii. "Being the person to whom Pat talked to"
      iii. "Being dangerous"
      iv. "Being trustworthy"
      v. … etc.





## Challenges (cont'd)

4. Connection
   a) What is "directly perceptible" is *that to which you are effectively (causally) connected*
   b) What **representation** is good for is to let you know *how things are to which you are not effectively connected* (for whatever reason)
5. Leads to a very important insight:



**Representation is that which allows you to "reach beyond your effective grasp"**

*"Ah, but a man's reach should exceed his grasp,*
*Or what's a heaven for?"*

*Sorry about the sexism …*

4. We can understand this in terms of what I call the "**REPRESENTATIONAL MANDATE**"





## "Representational Mandate" (I — Simple)

*i.e., that to which you can be "connected"*

1. Exploit what is **local and effective**

*where the normative constraints get a grip*

2. So as to *behave appropriately* with respect to (i.e., satisfy governing semantic norms regarding) that which is **distal and non-effective.**

## "Representational Mandate" (II — More Complex)

1. **Conditions**
   a) An intelligent system must work, *effectively*, in virtue of its *concrete material embodiment* ("materialism")
   b) Overall, it is *normatively directed* towards the **world as a whole**, including much that is *not effectively available* (distal, etc.)
   c) Being neither oracle nor angel, it has *no divine* (magical) *access* to those *non-effectively-available states that it cares about*.
2. **So what does it do?**
   a) It exploits *local, effective properties* that it **can use**, but does **not** (intrinsically) **care about**—including both:
      i. Interior (such as the internal configuration of its brain), and
      ii. Interactions with local, effectively available aspects of its environment
   b) To "stand in for" or "serve in place of" effective connection with states it cannot be effectively coupled to; in order to
   c) **Behave appropriately** towards those remote or distal (non-effective) states that it **does care about**, but **cannot use**.

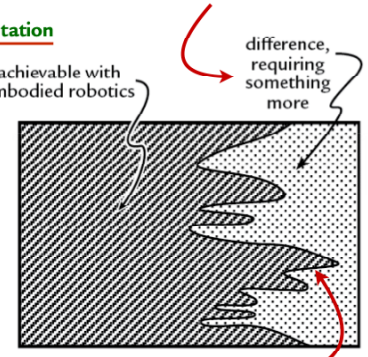
*The "extended mind," which we going to talk about next week*





**Q2:** ***What is this "more"?***

**A2**: The "something more" is **representation**




**The Challenge to Embodied Robotics**

**Q1:** ***What determines this boundary?***

**A1**: The **limits of effective reach** (i.e., limits of what the system can be *effectively (causally) connected to*

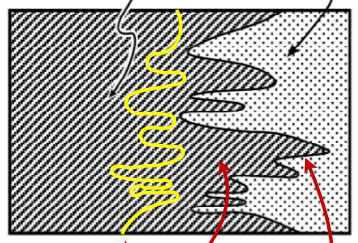




**Leads to the following general picture:**

Achievable via **direct engagement** (perception/action/coupling)

Requires **representation**



**The Challenge to Embodied Robotics** ~~of Mind~~

**the limits of the body**

**external scaffolding** — texts, computers, signs, etc.

the limits of **effective connection** / **causal reach**





**Cf. Karl Popper**

**"We think so that our hypotheses can die in our stead"**

*Tai!*

(= Think about it)

## Summary of the Course So Far

1. The **REPRESENTATIONAL MANDATE** provides an overarching principle in terms of which to design mechanical minds:
   a) Exploit what is **local and effective**, so as to *behave appropriately* with respect to the world as a whole, including *both*
      i. What is **locally and immediately (effectively) available**, and also
      ii. That which is **distal and non-effective**
   b) I.e., construct an effective mechanical system that is governed by *normative* (but non-effective) *constraints* on its flourishing
2. To the extent that you can do this by "using the world as its own model"—i.e., by direct perception, abstraction, and effective engagement with it, *do that!* It will be the most accurate and direct way to figure out what is going on.
3. To the extent that you *can't*, use a system of representation, in order to be able to coordinate appropriately with what is beyond your effective grasp.
4. Note: this is just the general "**CLASSICAL MODEL**" we talked about in Lecture 06 (b)!* ✔

*See slides 2-4 of Week 6b





## Summary of the Course So Far (cont'd)

5. We can now see that the **CLASSICAL MODEL** is not just (i) an abstraction in terms of which to understand GOFAI and logic-based AI, but (ii) a general theoretical perspective in terms of which to assess the merits and demerits of *all four mental architectures we have talked about*:
   **a) Good old-fashioned AI (GOFAI)**
   **b) Neurally-inspired machine networks (connectionism)**
   **c) Dynamical systems**
   **d) Behaviourally-based embodied, interactive robots**
6. It will also provide us with a theoretical perspective in terms of which to understand the next major topic we need to look at: the **extended mind**. ✔





## Summary of the Course So Far (cont'd)

7. In broad outline:
   a) GOFAI—Perhaps excessively rigid (because of the formality of its encodings), and no resources to explain where concepts come from, how genuine learning works, etc., but nevertheless the tradition has explored some very sophisticated and subtle issues at the *highly representational* end of the spectrum.
   b) The (Brooksian) embodied robotic tradition has explored the other end of the spectrum, and is most successful at the *directly engaged* "coupled" aspects of intelligent life.
   c) The neural-network (connectionist) tradition is spectacularly successful in exploring the domains of *perception*, and *recognition*—including conceptual recognition, where the flood of impinging signals are classified/catalogued in conceptual terms. We can imagine how it could be harnessed en route to conceptual representation (of the sort AI presumes), and its notion of concepts might, by being more fluid and imprecise, might provide a better basis for conceptual thought than GOFAI presumes. But, per se, it doesn't illuminate how such conceptual reasoning would go.
   d) And the dynamical approach provides us some "external" theoretical tools to get at some of the dynamical regularities of intelligent behaviour—specially appropriate, it would seem, for direct physical kinds of connection.





## Summary of the Course So Far (cont'd)

8. As for "extended mind", we can see the outline of what we will explore in more detail next week:
   a) A clear and clever strategy for intelligence is to populate our world with directly-engageable-with symbols and other external structures—what in general we will call "**scaffolding**"—to shoulder some of the representational load, with respect to the distal or otherwise inaccessible aspects of the world that we care about, so that we can directly engaged *with that external scaffolding*, instead of having to represent those things internally (in our own brain configurations).

**We will talk more about the extended mind on Tuesday (Nov 22).**

**In preparation, please Andy Clark and David Chalmers' "The Extended Mind"—(available on Blackboard)**

Today!

# Schedule

**Lecture C · 10** **Situated Cognition** (*Embodied*, *Embedded*, *Extended*, *Enactive Mind*)

## Part IV — Open Issues

**Lecture D · 01** **Consciousness I** — What Has Been Said (including *qualia*)
Read Nagel's "What is it Like to be a Bat?"

**Lecture D · 02** **Consciousness II** — What I Think (i.e., "the answer" ;))

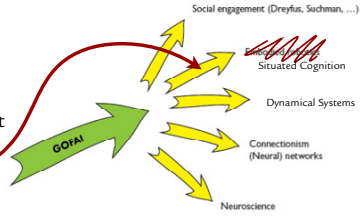
**Lecture D · 03** **Ethics of AI** — Lecture by Atoosa Kasirzadeh

**Lecture D · 04** **The Singularity** (and its Discontents)





## Situated Cognition




A. History
   a) In the 1980s, a "**situated cognition**" movement arose, in response to some of GOFAI's failures
   b) Brooks' "embodied robots" were one example (the one we looked at last week), but the movement was more general.
   c) The basic idea was that the *location*, *embodiment*, and *contextual situation* of an animal or system is of absolute importance *to understanding its mind*.

B. Importance
   a) No one denies (or ever would have denied, except perhaps dualists) that such systems *had* bodies, *were* physically located, *did* exist in contexts, etc.
   b) What people had thought, however, wrt logic in particular, but also wrt GOFAI, was that place, context, etc, were of *secondary theoretical importance* ("complicating incidents," as it were)—that the fundamentals of mind could be understood in ways that *abstracted away from those "complicating" contextual particulars* (much as it was thought that mind could be understood in ways that abstracted away from the details of its neurological implementation)
   c) The "situated movement" argued, on the contrary, that facts about body, place, time, context, etc., were **theoretically fundamental**—essential to what cognition is.





## Indexicality

1. Before we get to the various varieties of situated cognition that people have explored, consider one of the simplest phenomena that drove interest in it (especially in the 1980s): that of **indexicality**
2. Some simple example of **indexical expressions:** *I*, *here*, *now*, *today*, *to the right*, etc.
   a) There is something that *same* about all utterances or occurrences of these words
      — We don't need 4 billion entries in the dictionary: $I_1$, $I_2$, $I_3$, … etc., for what the word 'I' means for me, what it means for you, what it means for your grandmother, etc.
   b) Yet there is also something *different* about different utterances of them.
      — When you say "I am hungry," you report on a different person than I do, when I utter the same words.
      — Similarly, today, when I say "today," I refer to *today*, whereas tomorrow, when I say the same thing, I thereby refer to *tomorrow*.
      — Similarly, two people can each yell "I'm right! You're wrong!"— *without agreeing!*
3. These are facts that every school child knows, but they are impressive—and interesting.
4. Though technical vocabularies differ, though it is common to say that there is a
   a) Single **meaning** for each of these words, but
   b) A different **reference** or **interpretation**, depending on the **context of use**.





## Indexicality (cont'd)

4. Whatever one calls the two facets, it is clear that a competent user of a language *has to understand both*:
   a) What is the **same** (among different utterances or uses of them), and
   b) What is **different**
5. They also have to understand *how the different things are systematically related to the context*
   a) E.g., that different uses of '*I*' refer to the *speaker of the utterance*
   b) E.g., that different uses of '*today*' refer to *they day on which the utterance was* made
   c) … etc.
6. In a sense, learning the "*meaning*" of indexical terms (and phrases) involves learning something like "how the referent (or reference) is related to the context of use."
7. One of the first claims of the situated cognition movement was that something like **indexicality underlies a great deal of human understanding**. Cf. John Perry's famous story:

> *"I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch."*

## Indexicality (cont'd)

8. Perry's point is that there is something **essential** about recognizing oneself in a first-person, indexical way—something that is not equivalent to any other co-referring term (such as "the shopper with the torn sack", which indeed does refer to himself.
9. Similarly, consider the difference in the cognitive impact of the following two statements:
   a) "A meteor is going to strike
      i. At 43°, 39', 53.61" North and 79°, 23', 22.97" West, at
      ii. 432,736,992,935,821,846 seconds from the beginning of the universe!"
   b) "A meteor is going to strike *in this very room, in 10 minutes!*"
10. To the first, you might say: "Interesting".
    To the second: "*Let's get the hell out of here!*"
11. Similarly, which of the following is it more likely that your brain will do:
    a) Say to your arms: "Right arm, extend south!" ✘
    b) Say *to your right arm*: "Extend forwards!" ✔
12. By the same token, it is more likely that the signal from your stomach says (b) than (a):
    a) "*I'm hungry!*" ✘
    b) "*Hungry*!" ✔





## Indexicality (cont'd)

13. The point is that indexical representations are *much closer to what is received from*, and *much closer to what is required for*, immediate *bodily (physical) action and reaction*.
14. Next Tuesday, I will argue that something underlying this kind of indexicality is actually responsible for some of **the qualitative character of consciousness!**





## Other developments that built towards Situated Cognition

1. Indexicality was highlighted as a central issue in Barwise & Perry's *Situations and Attitudes* (1983), and was one topic that drove cognitive science towards a situated view of cognition.
2. Another impetus was provided by the 1987 publication of Lucy Suchman's *Plans and Situated Action*, and the research in her group at PARC, which focused on people's social, engaged forms of improvisational interaction with the environment, using methods and techniques from ethnomethodological anthropology and sociology.
3. Also published in 1987 was David Chapman and Phil Agre's "Pengi: an implementation of a theory of activity" (on Blackboard), with this abstract:

   > *AI has generally interpreted the organized nature of everyday activity in terms of plan-following. Nobody could doubt that people often make and follow plans. But the complexity, uncertainty, and immediacy of the real world require a central role for moment-to-moment improvisation. But before and beneath any planning ahead, one continually decides what to do now. Investigation of the dynamics of everyday routine activity reveals important regularities in the interaction of very simple machinery with its environment. We have used our dynamic theories to design a program, called Pengi, that engages in complex, apparently planful activity without requiring explicit models of the world.*

4. Note that Brooks' "Intelligence without Representation" was *also published in 1987!*
5. As these publications attest, the mid-1980s were a time when a "**sea change**" led us out of GOFAI into the varieties that we know today.





## Contemporary Varieties of Situated Cognition

1. Situated cognition evolved in such a way that one now hears people identity with, or endorse, four "kinds" or "flavours," each with its own emphasis
   a) **Embodied** · Cognition depends on facts about the *concrete, physical body*
   b) **Embedded** · Cognition arises in a system *embedded in a larger world*
   c) **Extended** · Cognition, not limited to the brain or body, *itself extends into the world*
   d) **Enactive** · Cognition depends on the *living body*, understood an *autonomous system, interacting* with its environment.
2. These aren't *alternatives* to the 4 architectures we've examined (GOFAI, plus three alternatives). Rather, in many ways, these four alternatives are basically *orthogonal to* (independent of) any specific architecture.
3. They should thus be viewed as **complementary** themes or perspectives
4. You can pledge allegiance to any one of themes themes—or more, or even all four—while retaining an architectural allegiance to any of the types we studied—or another one, or any combination.
5. We could (and should!) spend a week on each of these...but alas there is no time
6. So just a few brief remarks on each flavour—to give you a flavour, and so that you can explore them on your own.

## Variety #1 — Embodied Mind

1. Introduction
   a) What is it to say that the mind is "embodied"?
   b) Is it just to reject Descartes, and embrace some form of physicalism?
   c) No—that is not how it is normally understood
   d) It is taken to mean something more specific, and consequential, than that
2. Brain
   a) The first thing one might think of has to do with the *brain*
   b) GOFAI claimed that the mind/brain worked in terms of its *formal* properties
   c) I have already said that what computing (and logic) calls "formal" properties (the ones we signified with red arrows) are in reality causal properties
   d) Neural networks are also modelled (albeit at a higher level of abstraction) on how the brain works
   e) So in a way we have already taken on board the idea that the *mind/brain is physical*
3. Body
   a) The main thrust of the "embodied cognition" or "embodied mind" movement, however, doesn't have to do with the brain
   b) Rather, it takes as a central claim about *mind* or *cognition* or *intelligence is* that it arises **within a concrete, physical body**





## Variety #1 — Embodied Mind (cont'd)

4. Some examples
   a) The ability to *understand space* depends on one's capacities and activities of *moving around*
      i. Chimps deprived of movement *can't see*, even if their eyes function normally
      ii. The vestibular-ocular reflex, fundamental to human vision
   b) Non-conceptual content—meaning from *movement and action* (not abstract conception)
      i. "Space for a piano"
      ii. A footstep behind you (Evan's example of the location of an intruder)
   c) The division of labour between what we need to represent, and what we can do directly
      i. The morals we took from Brooks
      ii. Grush: representation in our arms, about the ballistics of reaching
   d) Lakoff and Johnson: "metaphors we live by"
      i. "Up" in hierarchies, judgment, idioms, etc.—based on *movement*, *bodily positioning*
      ii. "Forward and backwards"—for both space and time (note the Greek idea that we *back into the future*, because we can "see" the past, but not the future)





## Variety #1 — Embodied Mind (cont'd)

5. There are two versions of the "embodied mind" thesis, of differing strengths:
   a) **Weaker**
      i. The mind (or intelligence) is *in* the brain, but it *requires/depends on the body* to be a mind
      ii. Mind can thus only be *understood* in terms of the body
   b) **Stronger**
      i. The mind (or intelligence) is not (simply) in the brain, *but in the body as a whole*
      ii. So if I amputate part of your body (your leg, say), I have *damaged your mind*
6. The issue is whether the *brain/body* boundary, if there is such a thing, is the *boundary of the mind*—or whether that is not a theoretically interesting or coherent line to draw.





## Variety #2 — Embedded Mind

1. A natural counterpoint to saying that mind is **embodied** is to say that it is **embedded**
   a) Cf. Haugeland's "Mind Embodied *and Embedded*" (on Blackboard; emphasis added)
2. We have talked about the mind's constitutive relations to the surrounding (embedding) world throughout the course—especially with respect to *semantics* (blue arrows)
   a) Referring to or thinking about things is a relation to the embedding world
   b) Similarly, the reference (interpretation) of indexical expressions and thoughts
3. Another issue, also involving semantics, is called "**externalism**"—about whether even *meaning* extends into the world (a view that Dretske also holds)
   a) Cf. Putnam's example about the difference between "beech" and "elm": he doesn't know anything about how they differ, yet he is able to use them separately, and to know, for example, that a tree in his front yard is a beech, and not an elm.
   b) Putnam claims that he can use these terms to mean different things because he *relies on expertise held within the community of which he is a part*.

## Variety #2 — Embedded Mind (cont'd)

4. But there are other properties of embeddedness beyond semantics—such as discussions of how we construct **scaffolding** in the world, on which our thoughts and cognition rely.
   a) Signs, markers, cairns, blazes, etc.
   b) iPhones, etc.
   c) And perhaps the simplest and most powerful example of all: **language itself**!
5. Note that all of these examples of epistemic scaffolding are explicable in terms of the representational model (and the "representational mandate") that we talked about in the last class.
   a) We have no direct access to the facts that are important to know (that there is a curve coming up, or a stoplight; to the voice of our friend; to where exit #22 is, on the freeway).
   b) We can't represent it, either, because we don't know
   c) What the sign, or marker, or text, or iPhone does, is to allow us to perceive, directly, something that does represent the distal facts we care about, so that we can end up in an appropriate action-governing representational state.



## Variety #3 — Extended Mind

1. Stronger even than the embedded mind approach is what is called **extended mind**.
2. The idea is not just that mind *relies* on the external world, even necessarily.
3. Rather, the extended mind thesis claims that the mind *literally extends into the world*
   a) I.e., part of your mind is (or at least can be) **literally out in the "external" world**.
   b) Standard examples (of "epistemic actions")
      i. Rearranging tiles while playing Scrabble, to assist in finding good words
      ii. Performing mathematical calculation using pen and paper—or with a calculator
      iii. Even you cell phone, or a co-dependent partner ;-)
   c) Cf. the discussion of Inga and Otto, in Clark & Chalmers, where
      i. Inga has a good memory"
      ii. Otto doesn't have a good memory (perhaps from brain damage), but uses a notebook for all his memories.
4. According to the extended mind thesis, Otto's notebooks are *part of Otto's mind*.





## Variety #4 — Enactive Mind

1. The term "**enactive mind**" is primarily associated with Evan Thompson and his colleagues
2. The intuition is based in part on phenomenology, in part on Buddhism, and in part on theories of self-organizing complex (biological) systems—perhaps even a new synthesis of all three.
3. The basic thesis is that
   a) Thinking isn't what matters about mind
   b) Rather, intelligence doesn't just depend on, or arise from, but is in fact constituted by and in, **engaged, participatory interaction** with the embedding world.





## Morals

1. I have a considerable sympathy with all four of these proposals
   a) The intuitions on which they are based are important
   b) Their shift in focus is by and large *salutary*, as a corrective on the logicist/GOFAI approach
2. But they need to be understood in terms of (even if sometimes in distinction to) the issues we have been looking at all semester
   a) Issues about **representation**—in the general sense we have discussed it, of using effectively available resources to orient a mind or organism/system to that which is beyond effective reach (important, for example, in order to understand the role of scaffolding in the embedded mind approach)
   b) Capabilities for **categorization, classification, abstraction**, etc. (in all proposals)
   c) Questions about **disjunction, negation, modelling**, etc.
3. They are alternatives to strict logicism/GOFAI—that is true
4. But they are perhaps best thought of as territories onto which better theories of mind open out…