Artificial general intelligence

Computational theory of mind(s) and cognition

Antal Péter, Bolgár Bence

ComBine Lab

Artificial Intelligence group

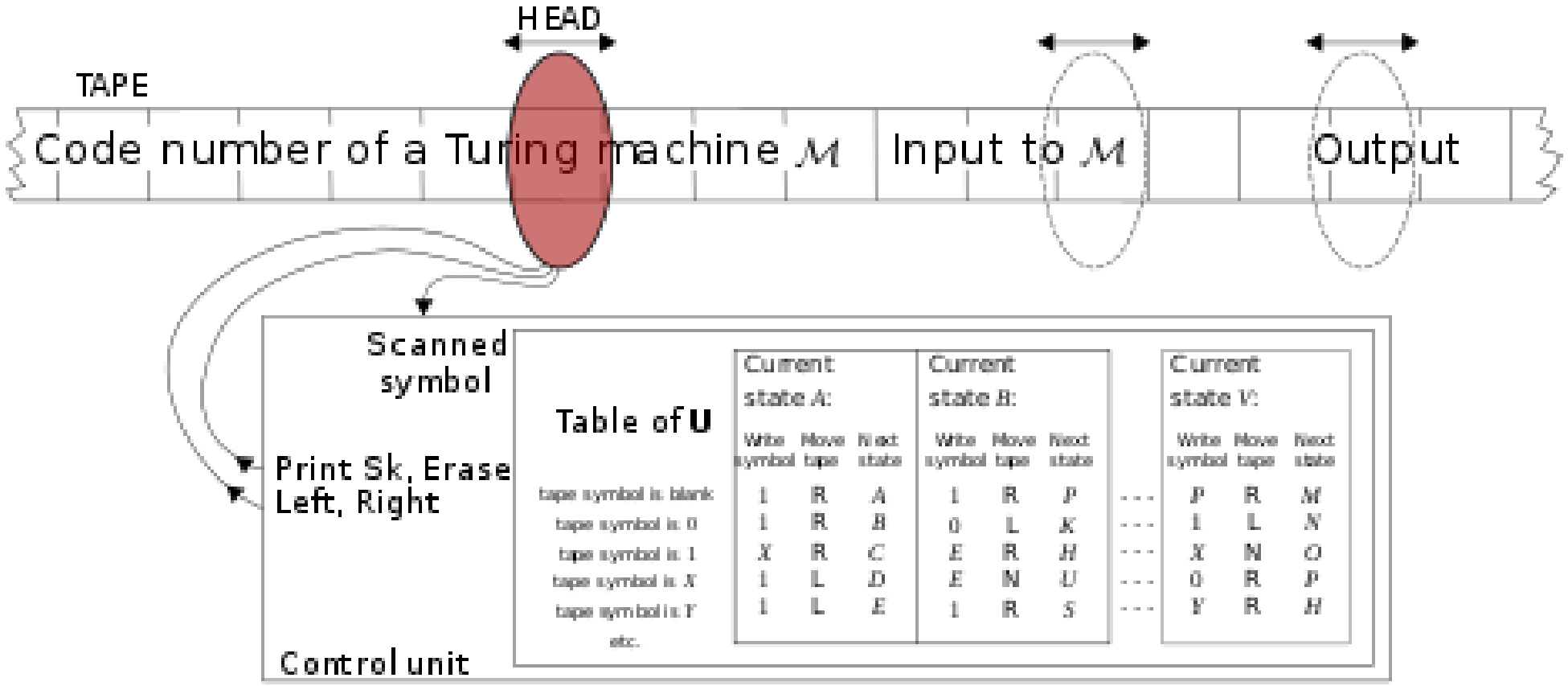
Department of Measurement and Information Systems

Agenda

- Easy(?) and hard problems around human-level intelligence
- Scientific building blocks
  - Universal (function) approximation theorems
  - Models of computation (architectures, complexities)
  - Computational linguistics: syntax and semantics
  - Quantum foundations
- Cognitive science
  - Architectures
  - Criticisms
- Next: the neural substrate

Extended computational models

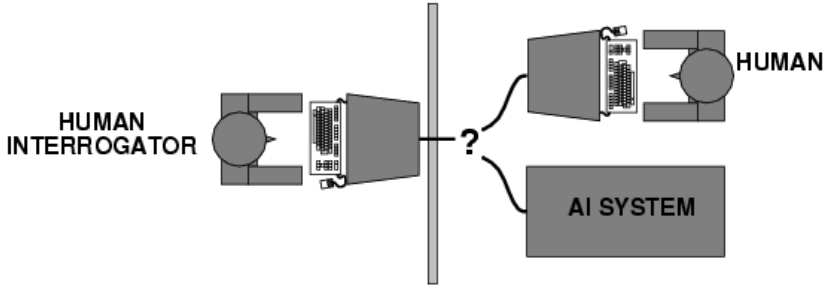
Universal Turing machine



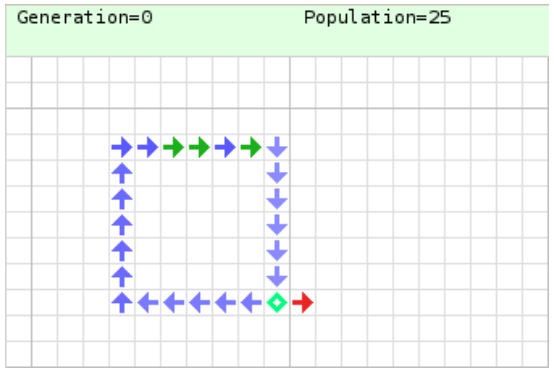
https://en.wikipedia.org/wiki/Universal_Turing_machine

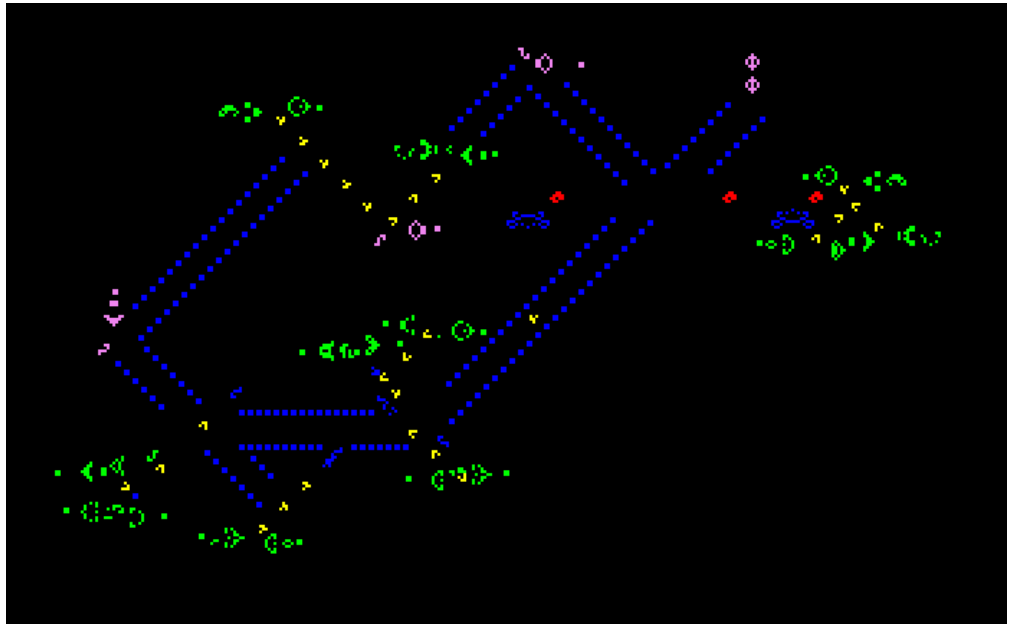
Turing Test

- Turing (1950) "Computing machinery and intelligence":
- "Can machines think?" → "Can machines behave intelligently?"
- Operational test for intelligent behavior: the Imitation Game



Universal Turing machine and cellular automaton





Extended computational models

- Computation over real numbers
  - chaotic systems in nature (neurons, brains)
  - Blum–Shub–Smale (BSS) machine
    - items: real numbers
    - operations: rational functions
- Non-deterministic machines
  - branching and parallel computation
- Probabilistic machines
  - stochastic operations
- Quantum machines
  - entangled states in superpositions

Easy and hard questions around AGI

Questions

- Interpretation of a compution
- Human bases of Turing test?
- Any non-(UTM)computable human feature?
- Any non-(UTM)computable intelligence feature?
- Any non-(UTM)computable, but well-defined intelligence feature?

Subjectivity and interpretation of computation

- Fred Hoyle: The black Cloud, 1957
- Implementation of a computation
  - John Searl's „wall": from a real-valued dynamical system any symbolic computation can be derived with proper mapping function
    - Similar arguments: Ian Hinckfuss, Hilary Putnam, David Chalmers
  - Complexity of the mapping function(!???)
    - Kolmogorov complexity
      - Ultimate compressibility of a string s: K(s)
      - Invariant ☺
      - Non-computable ☹
      - https://en.wikipedia.org/wiki/Kolmogorov_complexity
  - Robustness of the mapping function (non-chaotic)
    - for random perturbations
    - for interventions

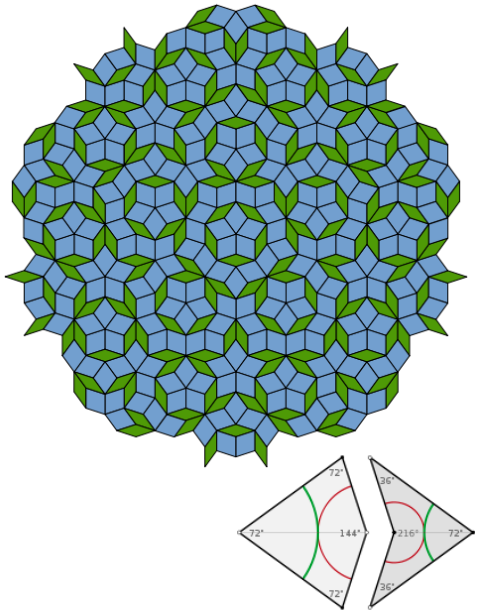
Non-computable human features?

- Movies
  - 2001: A Space Odyssey: +
  - Blade runner: ?
  - Interstellar: +
  - Her: -
  - …
- Extensions of Turing test:
  - The Coffee Test (Wozniak)
  - The Robot College Student Test (Goertzel)
  - The Employment Test (Nilsson)
  - The flat pack furniture test (Severyns)
  - …
- Creativity, scientific discovery
- Human values
- Human consciousness

„Holistic” constraints: aperiodic tiling

- A tessellation of the plane or of any other space is a cover of the space by closed shapes, called tiles, that have disjoint interiors.

- A Penrose tiling:
  - It is non-periodic (lacks any translational symmetry).
  - It is self-similar.
  - It is a quasicrystal (as a physical structure).



- How can we find such exotic „patterns”?
- R.Penrose: Emperor’s new mind

  https://en.wikipedia.org/wiki/Aperiodic_tiling
  https://en.wikipedia.org/wiki/Penrose_tiling

Non-computable, but intelligence features?

- Penrose–Lucas argument
  - https://en.wikipedia.org/wiki/Penrose%E2%80%93Lucas_argument
- Practical suggestions
  - heuristics, chunking, multi-aspects reasoning
  - diagrammatic/visual reasoning
  - natural language processing: understanding, translation, extraction, generation
  - modal logic, counterfactual reasoning
  - long-term/hierchical planning
  - naiv physics
  - learning: with background knowledge, multi-task learning, transfer learning,..
  - cooperation, self-organization(???)
- Values*: the value alignment problem
- Qualia*: the hard problem of consciousness

- *: necessary for/functional role in general intelligence? Human level management?

Scientific building blocks in the 50s for cognitive science

Universal approximation theorems

Kolmogorov (1957):

Theorem 2.3.1 (Kolmogorov, 1957): *Any continuous real-valued functions* $f(x_1, x_2, \ldots, x_n)$ *defined on* $[0, 1]_n$, $n \geq 2$, *can be represented in the form*

$$f(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{2n+1} g_j\left(\sum_{i=1}^{n} \phi_{ij}(x_i)\right) \quad (2.3.1)$$

where the g_j*'s are properly chosen continuous functions of one variable, and the* ϕ_{ij}*'s are continuous monotonically increasing functions independent of* f.

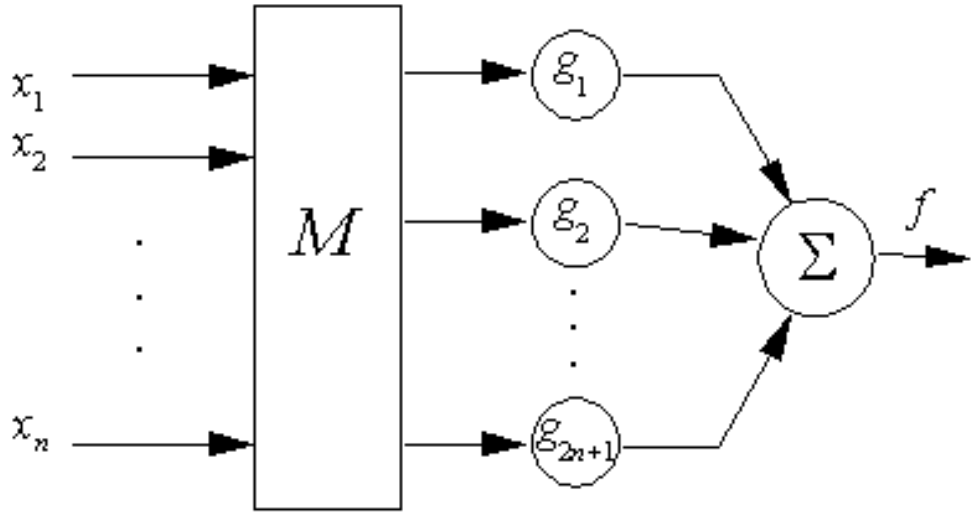


Figure 2.3.1. Network representation of Kolmogorov's theorem.

http://neuron.eng.wayne.edu/tarek/MITbook/chap2/2_3.html

Universal approximation theorems II.

- George Cybenko (1989): for sigmoid activation functions
- Width-bounded (depth-free) results:
  - Lu et al. (2017) width-n+4 networks with ReLU activation functions can approximate any Lebesgue integrable function on n-dimensional input space if network depth is allowed to grow.

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a nonconstant, bounded, and continuous function (called the *activation function*). Let I_m denote the m-dimensional unit hypercube $[0,1]^m$. The space of real-valued continuous functions on I_m is denoted by $C(I_m)$. Then, given any $\varepsilon > 0$ and any function $f \in C(I_m)$, there exist an integer N, real constants $v_i, b_i \in \mathbb{R}$ and real vectors $w_i \in \mathbb{R}^m$ for $i = 1, \ldots, N$, such that we may define:

$$F(x) = \sum_{i=1}^{N} v_i \varphi\left(w_i^T x + b_i\right)$$

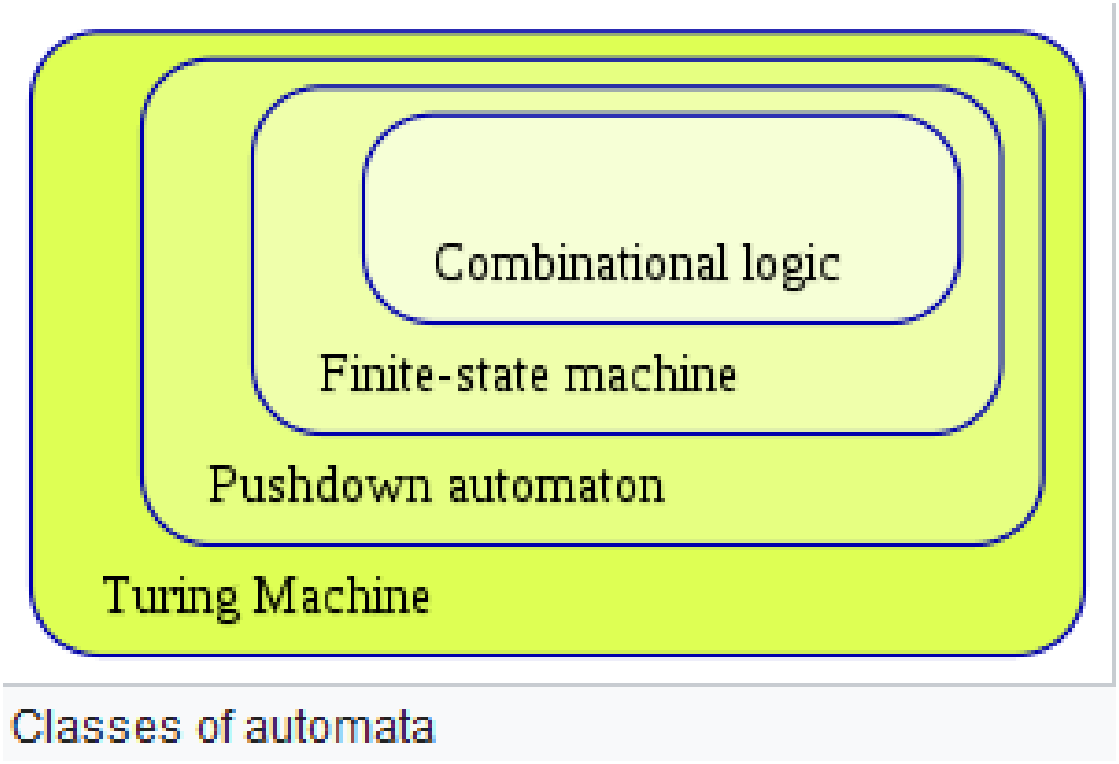
as an approximate realization of the function f; that is,

$$|F(x) - f(x)| < \varepsilon$$

for all $x \in I_m$. In other words, functions of the form $F(x)$ are dense in $C(I_m)$.

https://en.wikipedia.org/wiki/Universal_approximation_theorem

Computational models



Classes of automata

https://en.wikipedia.org/wiki/Automata_theory

Automaton
Deterministic Finite Automaton (DFA) -- Lowest Power
(same power) || (same power)
Nondeterministic Finite Automaton (NFA)
(above is weaker) ∩ (below is stronger)
Deterministic Push Down Automaton (DPDA-I)
with 1 push-down store
∩
Nondeterministic Push Down Automaton (NPDA-I)
with 1 push-down store
∩
Linear Bounded Automaton (LBA)
∩
Deterministic Push Down Automaton (DPDA-II)
with 2 push-down stores
||
Nondeterministic Push Down Automaton (NPDA-II)
with 2 push-down stores
||
Deterministic Turing Machine (DTM)
||
Nondeterministic Turing Machine (NTM)
||
Probabilistic Turing Machine (PTM)
||
Multitape Turing Machine (MTM)
||
Multidimensional Turing Machine

(Formal) grammars

N. Chomsky: separation of syntax and semantics

A grammar is formally defined as the tuple (N, Σ, P, S). Transformational grammar/rewrite system

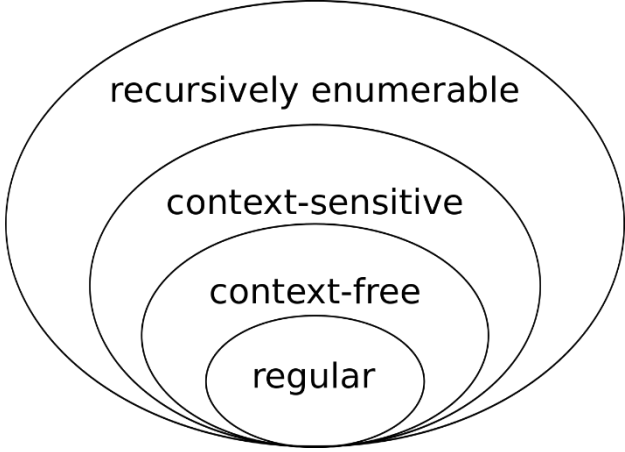
- A finite set *N* of *nonterminal symbols*, that is disjoint with the strings formed from *G*.
- A finite set Σ of *terminal symbols* that is disjoint from *N*.
- A finite set *P* of *production rules*, each rule of the form

$$(\Sigma \cup N)^* N (\Sigma \cup N)^* \rightarrow (\Sigma \cup N)^*$$

- A distinguished symbol $S \in N$ that is the *start symbol*, also called the *sentence symbol*.

https://en.wikipedia.org/wiki/Formal_grammar

The Chomsky hierarchy



Grammar	Languages	Automaton	Production rules (constraints)*	Examples[3]
Type-0	Recursively enumerable	Turing machine	$\alpha A \beta \to \gamma$	$L = \{w \mid w \text{ describes a terminating Turing machine}\}$
Type-1	Context-sensitive	Linear-bounded non-deterministic Turing machine	$\alpha A \beta \to \alpha \gamma \beta$	$L = \{a^n b^n c^n \mid n > 0\}$
Type-2	Context-free	Non-deterministic pushdown automaton	$A \to \alpha$	$L = \{a^n b^n \mid n > 0\}$
Type-3	Regular	Finite state automaton	$A \to \mathrm{a}$ and $A \to \mathrm{a}B$	$L = \{a^n \mid n \geq 0\}$

* Meaning of symbols:

- a = terminal
- A, B = non-terminal
- α, β, γ = string of terminals and/or non-terminals
  - α, β = maybe empty
  - γ = never empty

https://en.wikipedia.org/wiki/Chomsky_hierarchy

Logic in general

- Logics are formal languages for representing information such that conclusions can be drawn
- Syntax defines the sentences in the language
- Semantics define the "meaning" of sentences;

Propositional logic: syntax

- Propositional logic is the simplest logic
- The proposition symbols P_1, P_2 etc are sentences
  - If S is a sentence, $\neg S$ is a sentence (negation)
  - If S_1 and S_2 are sentences, $S_1 \wedge S_2$ is a sentence (conjunction)
  - If S_1 and S_2 are sentences, $S_1 \vee S_2$ is a sentence (disjunction)
  - If S_1 and S_2 are sentences, $S_1 \Rightarrow S_2$ is a sentence (implication)
  - If S_1 and S_2 are sentences, $S_1 \Leftrightarrow S_2$ is a sentence (biconditional)

Propositional logic: semantics

Assuming that true/false values are specified for each proposition symbol

E.g.

$P_{1,2}$	$P_{2,2}$	$P_{3,1}$
false	true	false

Rules (truth tables) for evaluating meaning (truth):

$\neg S$	is true iff	S is false	
$S_1 \wedge S_2$	is true iff	S_1 is true and	S_2 is true
$S_1 \vee S_2$	is true iff	$S_1$is true or	S_2 is true
$S_1 \Rightarrow S_2$	is true iff	S_1 is false or	S_2 is true
i.e.,	is false iff	S_1 is true and	S_2 is false
$S_1 \Leftrightarrow S_2$	is true iff	$S_1 \Rightarrow S_2$ is true and	$S_2 \Rightarrow S_1$ is true

On truth: entailment

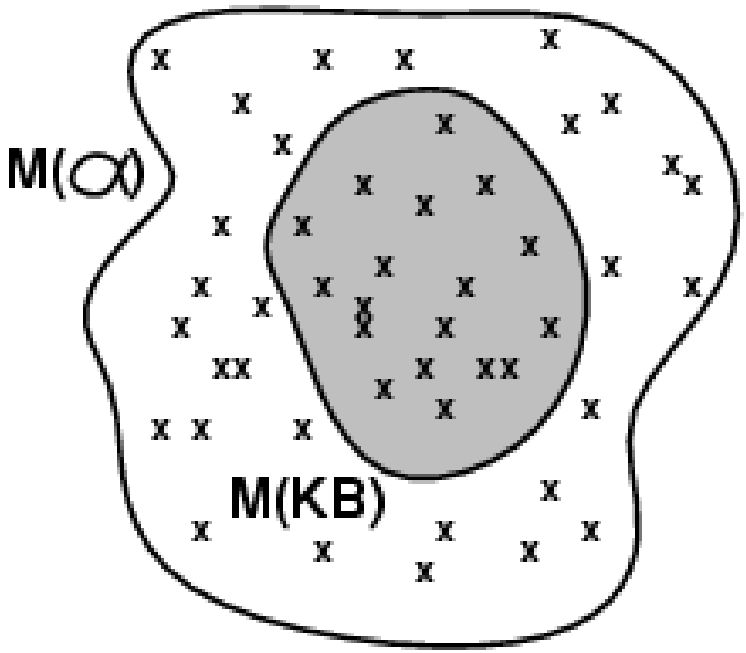
- Entailment means that one thing follows from another:

$$KB \models \alpha$$

- Knowledge base *KB* entails sentence α if and only if α is true in all worlds where *KB* is true.

Models

- Logicians typically think in terms of models, which are formally structured worlds with respect to which truth can be evaluated
-
- We say *m* is a model of a sentence α if α is true in *m*
- *M(α)* is the set of all models of α
-
- Then KB $\models$ α iff $M(KB) \subseteq M(\alpha)$
-



On proof

- $KB \vdash_i \alpha$ = sentence α can be derived from *KB* by procedure *i*
- Inference methods divide into (roughly) two kinds:
  - Application of inference rules
    - Legitimate (sound) generation of new sentences from old
    - Proof = a sequence of inference rule applications
      Can use inference rules as operators in a standard search algorithm
      - E.g. Modus Ponens, Modus Tollens, resolution
      - Typically require transformation of sentences into a normal form, e.g. into Conjunctive Normal Form (CNF)
  - Model checking
    - truth table enumeration (always exponential in *n*)
    - improved backtracking, e.g., Davis--Putnam-Logemann-Loveland
    - heuristic search in model space (sound but incomplete)
      e.g., min-conflicts-like hill-climbing algorithms

On truth and proof: $\models$?? $\vdash$

- Soundness: *i* is sound if whenever $KB \vdash_i \alpha$, it is also true that $KB \models \alpha$
- Completeness: *i* is complete if whenever $KB \models \alpha$, it is also true that $KB \vdash_i \alpha$

Quantum foundations

- Heisenberg's uncertainty principle
  - https://en.wikipedia.org/wiki/Uncertainty_principle
- Einstein-Podolsky-Rosen paradox: entangled states
  - https://en.wikipedia.org/wiki/EPR_paradox
- Schrödinger's cat: quantum state
  - https://en.wikipedia.org/wiki/Schr%C3%B6dinger%27s_cat
- Wigner's collapse: conscious observer can collapse the wave function in a quantum measurement
  - https://en.wikipedia.org/wiki/Von_Neumann%E2%80%93Wigner_interpretation

Computer models of mind/cognition

Books

- Boden, M., 1980. Artificial intelligence and natural man.
- Hofstadter, Douglas R. Gödel, Escher, Bach. Penguin Books, 1980.
- Boden, Margaret A. Computer models of mind. Cambridge University Press, 1988.
- Mérő, László. *Észjárások: a racionális gondolkodás korlátai és a mesterséges intelligencia*. Akadémiai Kiadó, 1989.
- Tibor, Vámos. Computer epistemology. Vol. 25. World Scientific, 1991.
- Pléh, Csaba. *A megismeréstudomány alapjai: az embertől a gépig és vissza*. Typotex, 2013.

Pléh Csaba: *A megismeréstudomány alapjai*

1. A MEGISMERÉSTUDOMÁNY (KOGNITÍV TUDOMÁNY) HELYE
2. A KOGNITÍV KUTATÁS KLASSZIKUS SZEMLÉLETE
3. A SZIMBÓLUMFELDOLGOZÓ GONDOLKODÁS NÉHÁNY RÉSZLETE
4. A SZIMBÓLUMFELDOLGOZÓ FELFOGÁS INHERENS BÍRÁLATA
5. A REPREZENTÁCIÓ FOGALMA A KOGNITÍV TUDOMÁNYBAN
6. A REPREZENTÁCIÓ „SZIGORÚBB” FOGALMA
7. GONDOLKODNAK-E A GÉPEK?
8. A KONNEKCIONISTA ALTERNATÍVA
9. A MODULOK PARLAMENTJE
10. BIOLÓGIAI ALTERNATÍVÁK
11. A TUDAT KÉRDÉSE A KOGNITÍV TUDOMÁNYBAN

A(G)I as “symbol manipulation”

- The Logic Theorist, 1955
  - ➔ see lectures on logic
- The Dartmouth conference ("birth of AI”, 1956)
- List processing (Information Processing Language, IPL)
- Means-ends analysis ("reasoning as search")
  - ➔ see lectures on planning
- The General Problem Solver
- Heuristics to limit the search space
  - ➔ see lecture on informed search
- The physical symbol systems hypothesis
  - intelligent behavior can be reduced to/emulated by symbol manipulation
- The unified theory of cognition (1990, cognitive architectures: Soar, ACT-R)

- Newel&Simon: Computer science as empirical inquiry: symbols and search, 1975

Constraints on mind

1. Behave as an (almost) arbitrary function of the environment (universality).
2. Operate in real time.
3. Exhibit rational, i.e., effective adaptive behavior.
4. Use vast amounts of knowledge about the environment.
5. Behave robustly in the face of error, the unexpected, and the unknown.
6. Use symbols (and abstractions).
7. Use (natural) language.
8. Exhibit self-awareness and a sense of self.
9. Learn from its environment.
10. Acquire its capabilities through development.
11. Arise through evolution.
12. Be realizable within the brain as a physical system.
13. Be realizable as a physical system.

Newell, A., 1980. Physical symbol systems. *Cognitive science*, *4*(2), pp.135-183.

A physical symbol system

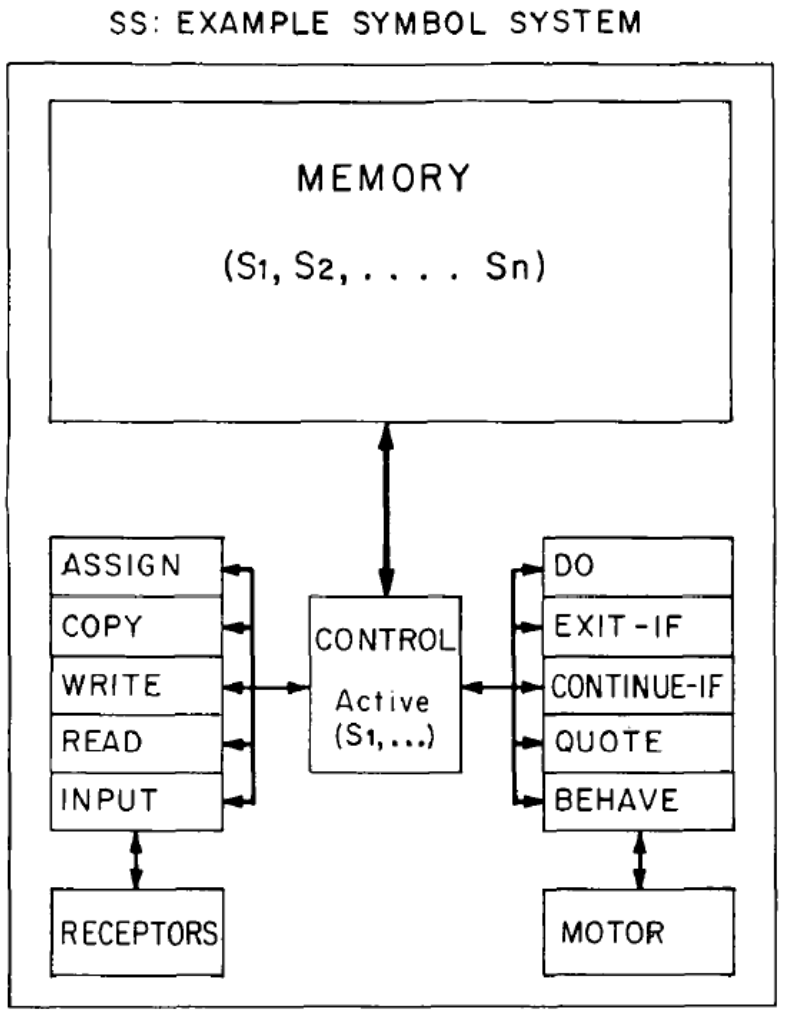


Figure 2. Structure of SS, a Paradigmatic Symbol System.

Newell, A., 1980. Physical symbol systems. *Cognitive science*, *4*(2), pp.135-183.

The physical symbol system hypothesis

Physical Symbol System Hypothesis: The necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system.

Necessary means that any physical system that exhibits general intelligence will be an instance of a physical symbol system.

Sufficient means that any physical symbol system can be organized further to exhibit general intelligent action.

General intelligent action means the same scope of intelligence seen in human action: that in real situations behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some physical limits.

Newell, A., 1980. Physical symbol systems. *Cognitive science*, *4*(2), pp.135-183.

Architectures: cognition

- SOAR
  - Newell, A., 1980. Physical symbol systems. *Cognitive science*, *4*(2), pp.135-183.
  - Laird, J.E., Newell, A. and Rosenbloom, P.S., 1987. Soar: An architecture for general intelligence. *Artificial intelligence*, *33*(1), pp.1-64.
  - Rosenbloom, P.S., Laird, J. and Newell, A. eds., 1993. The SOAR papers: Research on integrated intelligence.
- ACT-R (Adaptive Character of Thought, ACT-R)
  - Anderson, J.R. and Bellezza, F.S., 1993. Rules of the mind. Hillsdale, NJ: L.
  - Anderson, J.R., 2014. *Rules of the mind*. Psychology Press.
  - Anderson, J.R., 1996. ACT: A simple theory of complex cognition. *American psychologist*, *51*(4), p.355.
  - Lebiere, C. and Anderson, J.R., 1993, June. A connectionist implementation of the ACT-R production system. In *Proceedings of the fifteenth annual conference of the Cognitive Science Society* (pp. 635-640).
  - Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C. and Qin, Y., 2004. An integrated theory of the mind. *Psychological review*, *111*(4), p.1036.
  - Anderson, J.R., 2009. *How can the human mind occur in the physical universe?* (Vol. 3). Oxford University Press.

http://act-r.psy.cmu.edu/

https://en.wikipedia.org/wiki/ACT-R

ACT-R

- a cognitive architecture
- a theory for simulating and understanding human cognition

ACT-R Theory
- Architecture

Language Processing
- Analogy and Metaphor
- Language Learning
- Lexical and General Language Processing
- Parsing
- Sentence Memory

Perception and Attention
- Attention
- Driving and Flying Behavior
- Eye Movements
- Graphical User Interfaces
- Multi-Tasking
- Psychophysical Judgements
- Situational Awareness and Embedded Cognition
- Stroop
- Subitizing
- Task Switching
- Time Perception
- Visual Search

Problem Solving and Decision Making
- Choice and Strategy Selection
- Dynamic Systems
- Errors
- Game Playing
- Insight and Scientific Discovery
- Mathematical Problem Solving
- Programming
- Reasoning
- Spatial Reasoning and Navigation
- Tower of Hanoi
- Use and Design of Artifacts

Learning and Memory
- Category Learning
- Causal Learning
- Cognitive Arithmetic
- Declarative Memory
- Implicit Learning
- Interference
- Learning by Exploration and Demonstration
- List Memory
- Practice and Retention
- Reinforcement Learning
- Representation
- Skill Acquisition
- Updating Memory and Prospective Memory
- Working Memory

Other
- Cognitive Development
- Cognitive Workload
- Communication, Negotiation, and Group Decision Making
- Comparative (Architectures)
- Comparative (Inter-species)
- Computer Generated Forces, Video Games, and Agents
- fMRI
- Individual Differences
- Information Search
- Instructional Materials
- Intelligent Tutoring Systems
- Motivation, Emotion, Cognitive Moderators, & Performance
- Neuropsychology
- Tools
- Unrelated to ACT-R
- User Modeling

Uncategorized

http://act-r.psy.cmu.edu/publication/

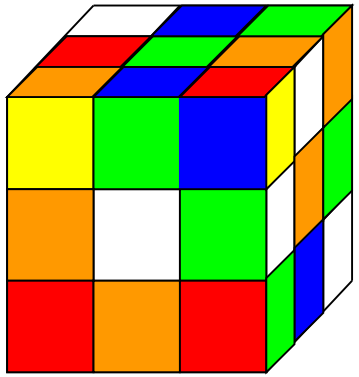
Nature of expertise
(in rule-based production systems)

- Complex symbols (schemata, gestalt, patterns)
  - Sub-symbolic learning
- Complex rules
  - Meta-learning
- **Efficient inference: heuristics** (sub-thinking the right thing ;-)
  - [dictionary]*"A rule of thumb, simplification, or educated guess that reduces or limits the search for solutions in domains that are difficult and poorly understood."*
  - *Prioritization of rules: timing, scoring,..*

Reminder: main properties of uninformed search

Criterion	Breadth-First	Uniform-cost	Depth-First	Depth-limited	Iterative deepening	Bidirectional search
Complete?	YES*	YES*	NO	YES, if $l \geq d$	YES	YES*
Time	b^{d+1}	$b^{C^*/e}$	b^m	b^l	b^d	$b^{d/2}$
Space	b^{d+1}	$b^{C^*/e}$	bm	bl	bd	$b^{d/2}$
Optimal?	YES*	YES*	NO	NO	YES	YES

Rubik's Cube



- The cardinality: 10^{19}
- Any position can be solved in 20 or fewer moves (where a half-twist is counted as a single move)
- average branching factor is ~13.3
- Invented in 1974 by Ernő Rubik.
- Rubik's cube current world records
  - http://www.youtube.com/watch?v=oC0B4b4J9Ys
- How can we guide the search process???

Agostinelli, F., McAleer, S., Shmakov, A. and Baldi, P., 2019. Solving the Rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, *1*(8), pp.356-363.

Inventing admissible heuristics

- Admissible heuristics can be derived from the exact solution cost of a relaxed version of the problem.

The optimal solution cost of a relaxed problem is no greater than the optimal solution cost of the real problem.

Another way to find an admissible heuristic is through learning from experience:

- Experience = solving lots of 8-puzzles
- An inductive learning algorithm can be used to predict costs for other states that arise during search.

`ABSolver` found a useful heuristic for the *Rubic cube*.

Prieditis: Machine Discovery of Effective Admissible Heuristics, 1993

Solving the Rubik's cube with deep reinforcement learning and search

The Rubik's cube is a prototypical combinatorial puzzle that has a large state space with a single goal state. The goal state is unlikely to be accessed using sequences of randomly generated moves, posing unique challenges for machine learning. We solve the Rubik's cube with DeepCubeA, a deep reinforcement learning approach that learns how to solve increasingly difficult states in reverse from the goal state without any specific domain knowledge. DeepCubeA solves 100% of all test configurations, finding a shortest path to the goal state 60.3% of the time. DeepCubeA generalizes to other combinatorial puzzles and is able to solve the 15 puzzle, 24 puzzle, 35 puzzle, 48 puzzle, Lights Out and Sokoban, finding a shortest path in the majority of verifiable cases.

Agostinelli, F., McAleer, S., Shmakov, A. and Baldi, P., 2019. Solving the Rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, *1*(8), pp.356-363.

Architectures: maps

- Spatial, cognitive maps
- Self-organizing map (SOM)
  - https://en.wikipedia.org/wiki/Self-organizing_map
- Thousand Brains Theory of Intelligence
  - https://numenta.com/blog/2019/01/16/the-thousand-brains-theory-of-intelligence/

Architectures: language

- Famous experiments about an inherent (~oldest) language
- N. Chomsky:
  - Universal grammar
  - Inherent language theory
- J. Fodor
  - Innate language module
  - Fodor, J.A., 1983. The modularity of mind. MIT press.
- S. Pinker
  - innate capacity for language
  - Pinker, S., 2003. The language instinct: How the mind creates language. Penguin UK.

Architectures: vision

- David Marr's Tri-Level Hypothesis:
  - computational level:
    - what does the system do (e.g.: what problems does it solve or overcome) and similarly, why does it do these things
  - algorithmic level (sometimes representational level):
    - how does the system do what it does, specifically, what representations does it use and what processes does it employ to build and manipulate the representations
  - implementational/physical level:
    - how is the system physically realised (in the case of biological vision, what neural structures and neuronal activities implement the visual system)

Beyond/below symbolic cognition

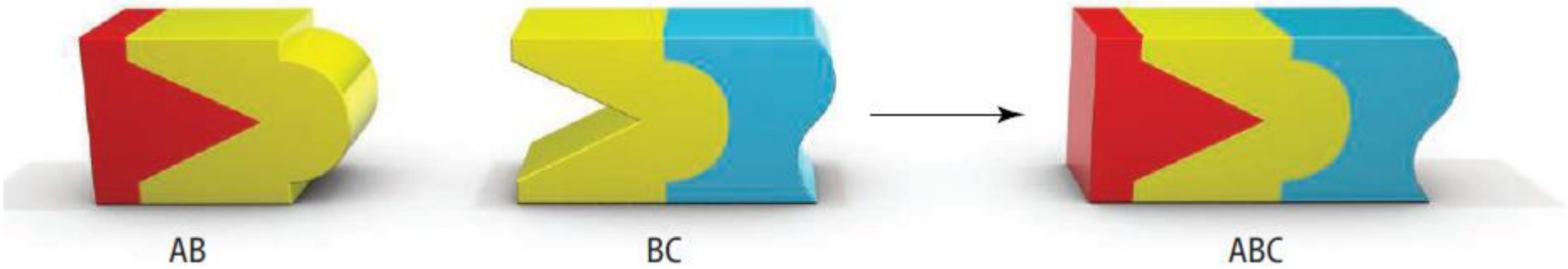
Collective cognition

- Kuhn, T.S., 1962. The structure of scientific revolutions. *Chicago and London*.
- Popper, K.R., 1972. *Objective knowledge* (Vol. 360). Oxford: Oxford University Press.
- Rosenberg, A., 2000. *Philosophy of science: A contemporary introduction*. Routledge.
- Rosenberg, A., 2018. *How History Gets Things Wrong: The Neuroscience of Our Addiction to Stories*. MIT Press.
- …

- Langley, P. (**1978**). Bacon: A general discovery system. Proceedings of the Second Biennial Conference of the Canadian Society for Computational Studies of Intelligence (pp. 173-180). Toronto, Ontario.
- …
- Chrisman, L., Langley, P., & Bay, S. (**2003**). Incorporating biological knowledge into evaluation of causal regulatory hypotheses. Proceedings of the Pacific Symposium on Biocomputing (pp. 128-139). Lihue, Hawaii.

Machine science

- Swanson, Don R. "Fish oil, Raynaud's syndrome, and undiscovered public knowledge." *Perspectives in biology and medicine* 30.1 (1986): 7-18.
- Smalheiser, Neil R., and Don R. Swanson. "Using **ARROWSMITH**: a computer-assisted approach to formulating and assessing scientific hypotheses." *Computer methods and programs in biomedicine* 57.3 (1998): 149-153.
- D. R. Swanson et al.: **An interactive system for finding complementary literatures: a stimulus to scientific discovery**, Artificial Intelligence, 1997

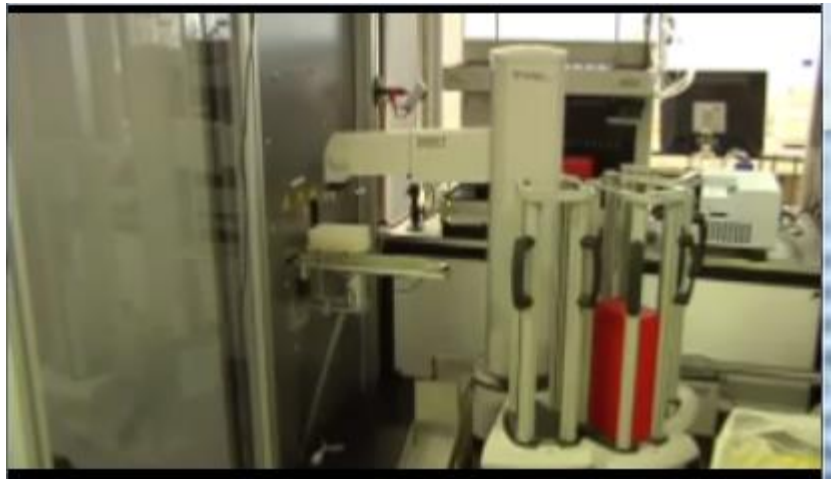


- James Evans and Andrey Rzhetsky: **Machine science**, Science, 2013

„Soon, computers could generate many useful hypotheses with little help from humans."

Automated discovery systems

- Yoo, C. and Cooper, G.F., 2004. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. Artificial Intelligence in Medicine, 31(2), pp.169-182.
- R.D.King et al.: The Automation of Science, Science, 2009
- Rzhetsky, A., Foster, J.G., Foster, I.T. and Evans, J.A., 2015. Choosing experiments to accelerate collective discovery. Proceedings of the National Academy of Sciences, 112(47), pp.14569-14574.



Values, utilities, utilitarianism

Theorem (Ramsey, 1931; von Neumann and Morgenstern, 1944):
Given preferences satisfying the constraints
there exists a real-valued function U such that

$$U(A) \geq U(B) \quad \Leftrightarrow \quad A \succsim B$$
$$U([p_1, S_1; \ \ldots \ ; \ p_n, S_n]) = \sum_i p_i U(S_i)$$

MEU principle:

Choose the action that maximizes expected utility

Note: an agent can be entirely rational (consistent with MEU)
without ever representing or manipulating utilities and probabilities

E.g., a lookup table for perfect tictactoe

https://en.wikipedia.org/wiki/Utilitarianism

Epiphenomenalism vs. causal power

- Principles for a causal relation between X➔Y:
  - **Probabilistic association**,
  - **Temporal asymmetry**: X precedes temporally Y,
  - ~~(Physical locality)~~
  - Quantitative effect of interventions: **dose-effect relation**
    - necessity (i.e., if the cause is removed, effect is decreased)
    - sufficiency (if exposure to cause is increased, effect is increased)
  - **Counterfactuals**:
    - Y would not have been occurred with that much probability if Y hadn't been present
    - Y would have been occurred with larger probability if X had been present
  - **Bounded context-sensitivity** (~context-free): relevant on average
  - Plausible **explanation** (no alternative based on confounding).

- Duality principle: rules/mechanisms vs. observations/interventions.

https://en.wikipedia.org/wiki/Epiphenomenalism

A SZIMBÓLUMFELDOLGOZÓ FELFOGÁS BÍRÁLATA

- *Nem komputáció*
- *Nem kategorizáció*
- *Nem propozíciók hálója*
- *Nem szimbólummanipuláció*
- *Nem pusztán reprezentáció*
- *Nem közegfüggetlen*
- *Nem statikus*

4.1 táblázat Mi mindent helytelenítenek az új felfogások a klasszikus kognitivizmusból

KLASSZIKUS SZEMLÉLET	*MEGKÉRDŐJELEZÉS ÉS FINOMÍTÁS*
egységes	moduláris
szimbolikus	szubszimbolikus
propozicionális	hálózatelvű
szekvenciális	párhuzamos
atomisztikus	procedurális
explicit	implicit
logikus, deduktív	intuitív, élményelvű
egyéni	szociális
testetlen	testre vonatkozó
önmagában tekinthető	evolúciós
modellálható	kimeríthetetlen
gépies, automatikus	emberi, jelentésorientált
igazságorientált	vágy irányította
tudásfüggetlen	tudás áthatotta

4.2 táblázat A klasszikus kognitivizmus és az alternatív irányok jellegzetes szembenállásai

Pléh Csaba, 2013. *A megismeréstudomány alapjai: az embertől a gépig és vissza*. Typotex.

Connectionism

- The neural „substrate”/architecture
  - Real-valued
    - + chaotic
  - Stochastic
  - ….

Summary

- The physical symbol system hypothesis of general intelligence
- Architectures of cognition
- Beyond/below symbols
  - **Heuristics, human biases**
  - Values, utilities
  - Causation
  - Implementation/Realization➔**Connectionism/neural networks**