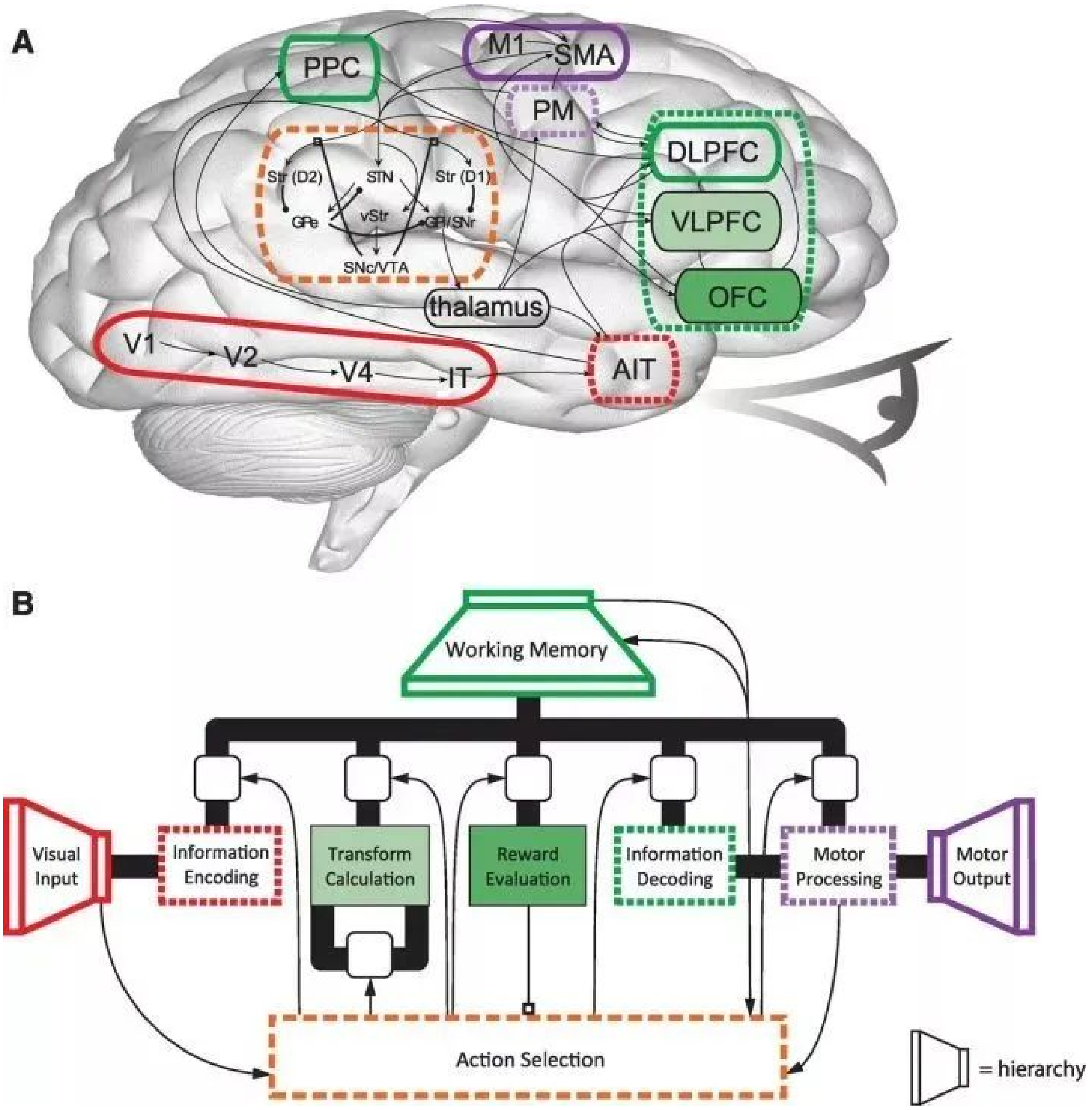
来源：人工智能前沿讲习

一、基于生物和经验的模型

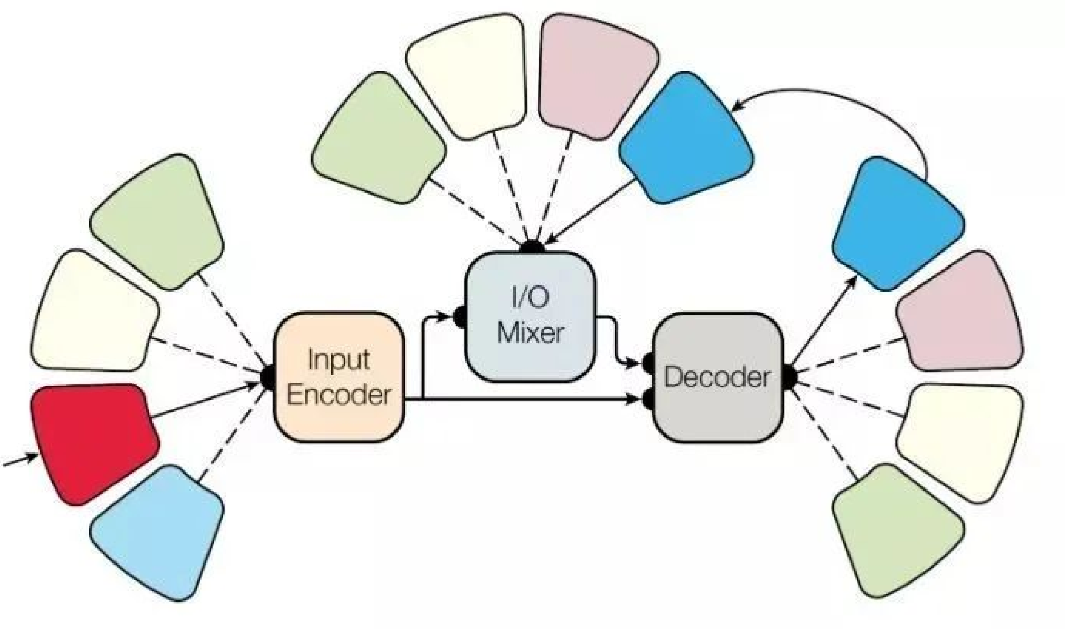
首先是 2012 年的 Spaun，基于生物基础（脑图谱），类生物神经元（尖峰放电 SNN）。
在训练后可完成多种识别和生成和反应任务。



1. map the visual hierarchy firing pattern to a conceptual firing pattern as needed
2. extract relations between input elements (transformation calculation)
3. evaluate the reward associated with the input (reward evaluation)
4. decompress firing patterns from memory to conceptual firing pattern (information decoding)
5. map conceptual firing patterns to motor firing patterns and control motor timing (motor processing)

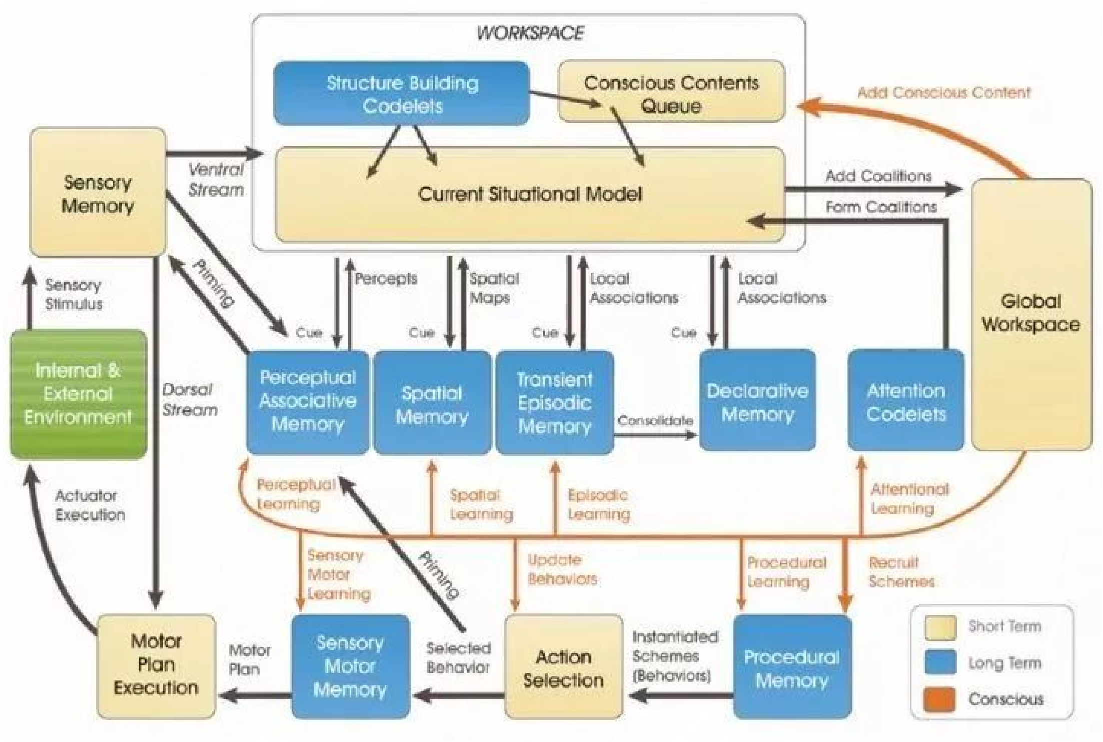
PPC, posterior parietal cortex; M1, primary motor cortex; SMA, supplementary motor area; PM, premotor cortex; VLPFC, ventrolateral prefrontal cortex; OFC, orbitofrontal cortex; AIT, anterior inferior temporal cortex; Str, striatum; vStr, ventral striatum; STN, subthalamic nucleus; GPe, globus pallidus externus; GPi, globus pallidus internus; SNr, substantia nigra pars reticulata; SNc, substantia nigra pars compacta; VTA, ventral tegmental area; V2, secondary visual cortex; V4, extrastriate visual cortex.

在许多深度学习模型中有类似的【编码-转换-解码】结构，例如 2017 年的 MultiModel ：



目前流行的大脑/意识理论是 GWT（Global Workspace Theory）和 IIT（Integrated Information Theory）。

GWT 用意识的功能描述意识：意识具有某些功能，例如输入输出和各种模块。典型的例子如下：



这种模型有悠久的历史和诸多例子：

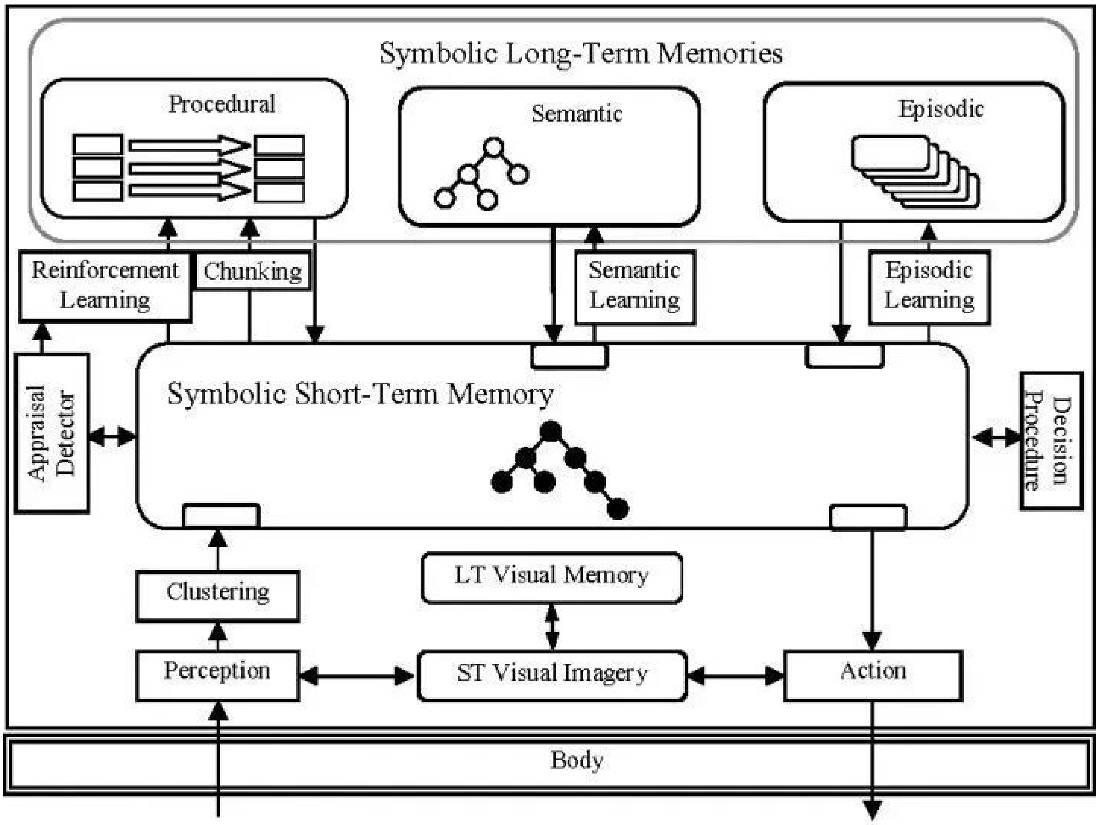
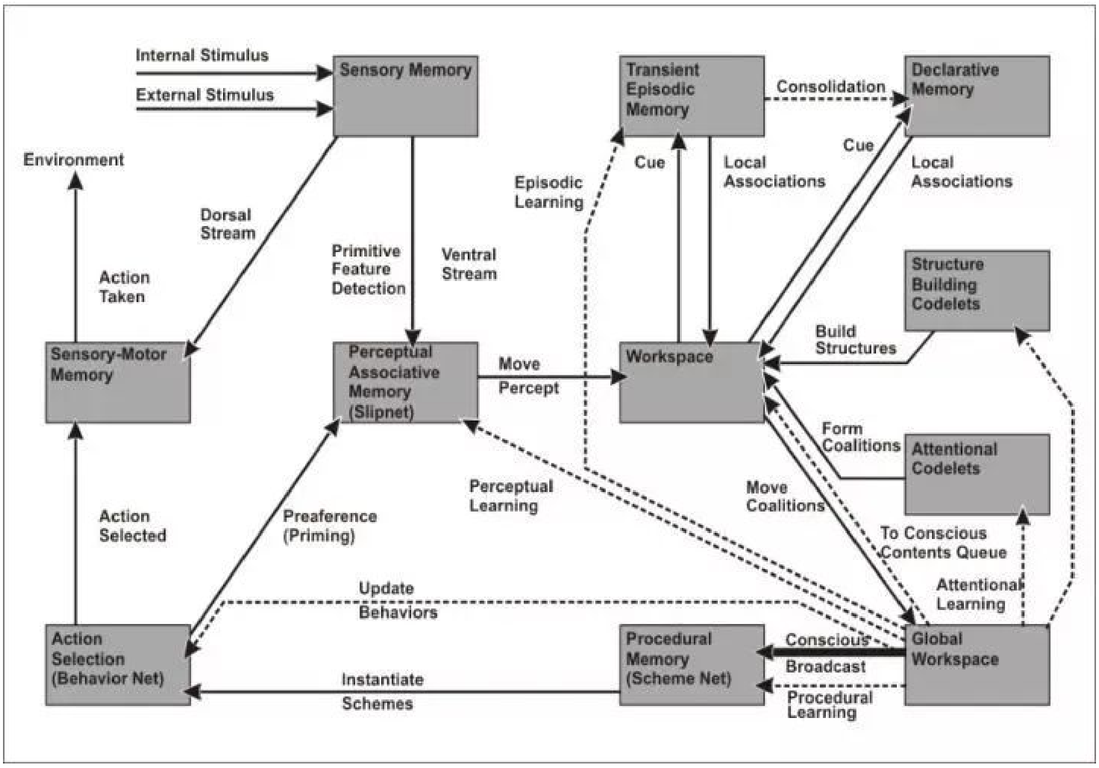
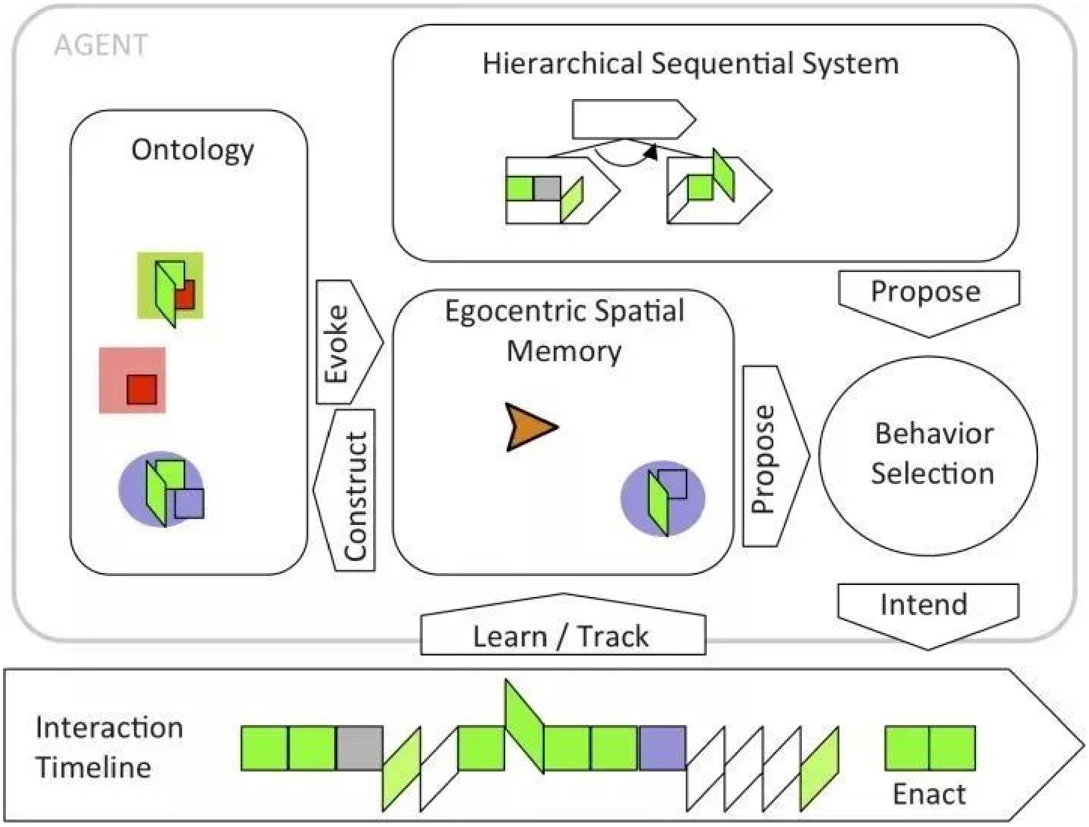
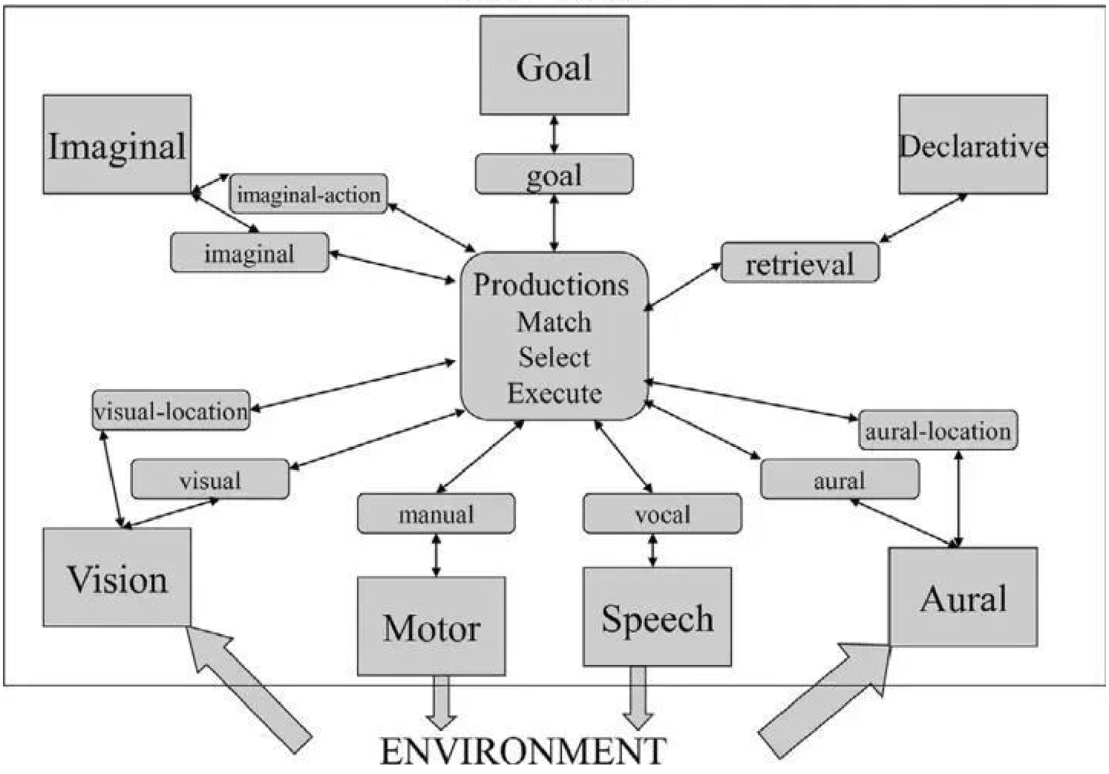


Figure 3: Soar 9

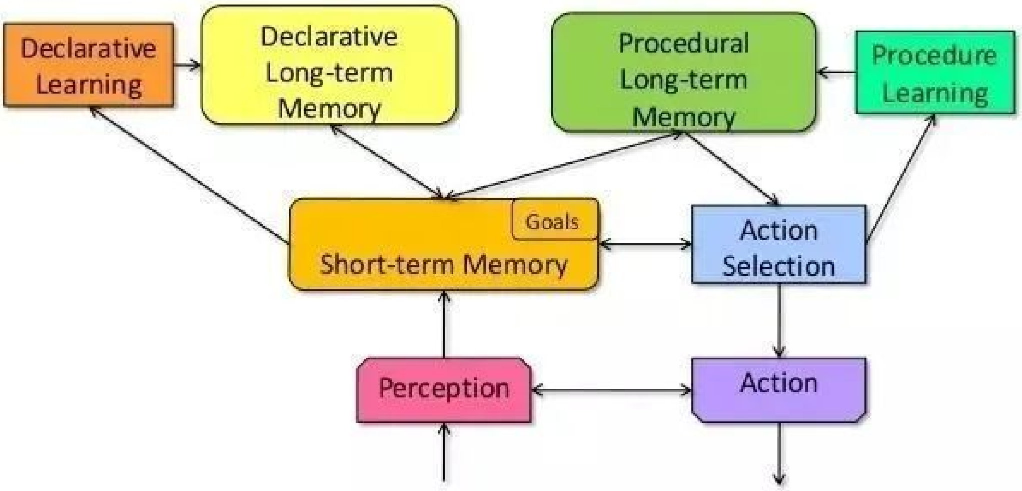

Internal Stimulus
External Stimulus
Sensory Memory
Environment
Dorsal Stream
Action Taken
Primitive Feature Detection
Ventral Stream
Sensory-Motor Memory
Perceptual Associative Memory (Slipnet)
Move Percept
Transient Episodic Memory
Consolidation
Declarative Memory
Episodic Learning
Cue
Local Associations
Cue
Local Associations
Structure Building Codelets
Build Structures
Workspace
Form Coalitions
Attentional Codelets
Action Selected
Preaference (Priming)
Perceptual Learning
Move Coalitions
To Conscious Contents Queue
Attentional Learning
Update Behaviors
Action Selection (Behavior Net)
Procedural Memory (Scheme Net)
Conscious Broadcast
Global Workspace
Instantiate Schemes
Procedural Learning


AGENT
Hierarchical Sequential System
Ontology
Evoke
Construct
Egocentric Spatial Memory
Propose
Propose
Behavior Selection
Intend
Learn / Track
Interaction Timeline
Enact

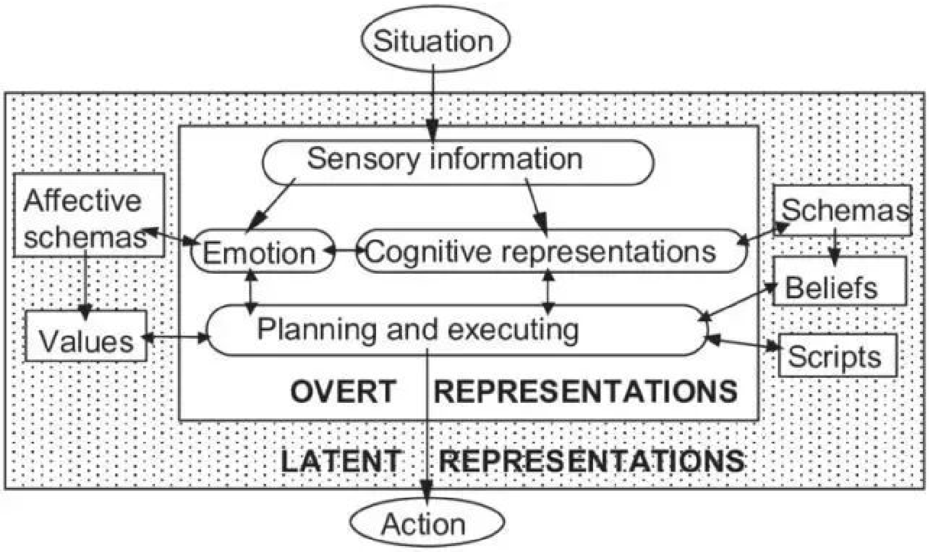
ACT-R 6.0

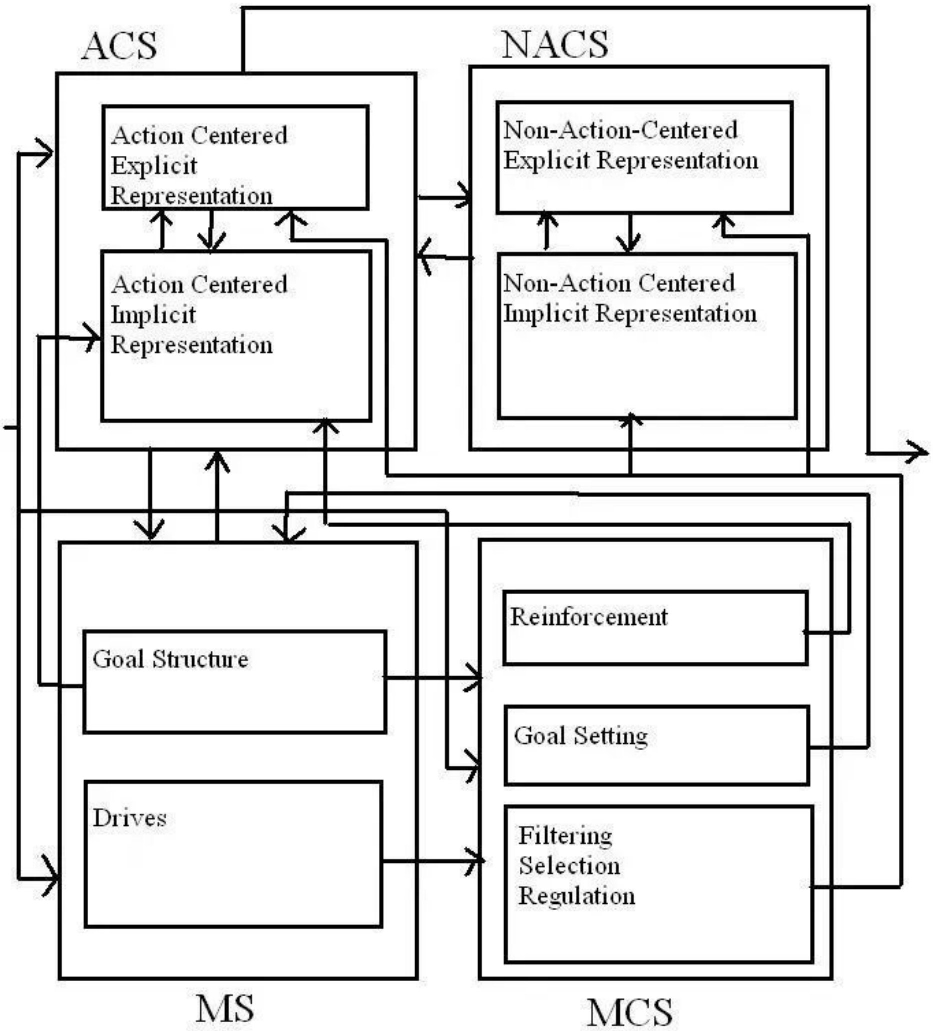


Common Structures of many Cognitive Architectures



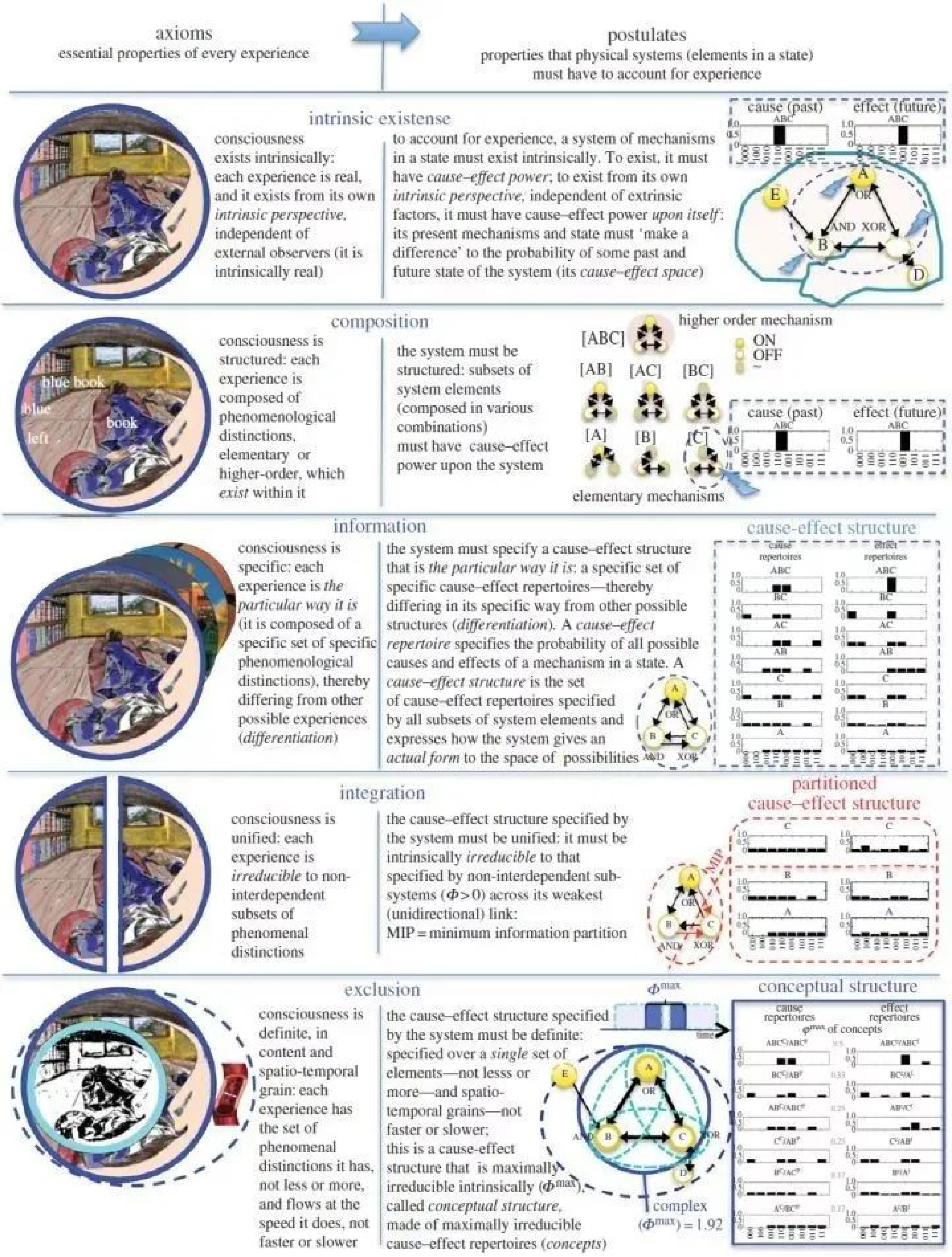
UNIVERSITY OF MICHIGAN




ACS
NACS
Action Centered Explicit Representation
Action Centered Implicit Representation
Non-Action-Centered Explicit Representation
Non-Action Centered Implicit Representation
Goal Structure
Drives
Reinforcement
Goal Setting
Filtering Selection Regulation
MS
MCS

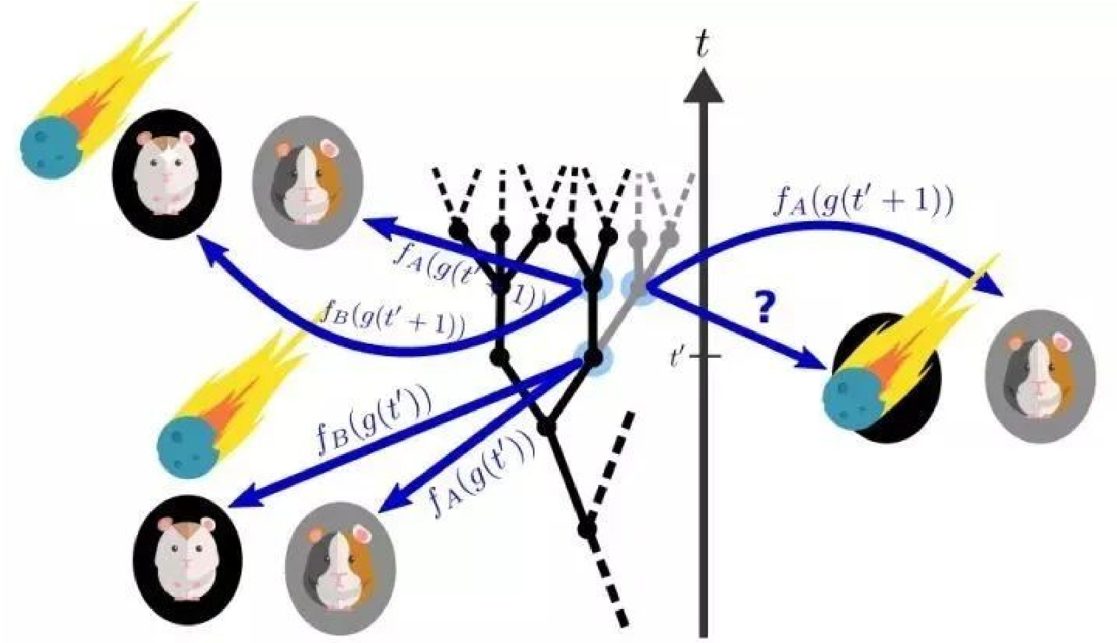
Clarion Cognitive Architecture

IIT 用意识的特征描述意识：意识是满足某些特征的现象。
虽然牺牲了一定的具体性，但也许更能同时描述非地球非碳基生物的意识。典型的特征如下：


axioms
essential properties of every experience
postulates
properties that physical systems (elements in a state) must have to account for experience
intrinsic existense
consciousness exists intrinsically: each experience is real, and it exists from its own intrinsic perspective, independent of external observers (it is intrinsically real)
to account for experience, a system of mechanisms in a state must exist intrinsically. To exist, it must have cause–effect power; to exist from its own intrinsic perspective, independent of extrinsic factors, it must have cause–effect power upon itself: its present mechanisms and state must 'make a difference' to the probability of some past and future state of the system (its cause–effect space)
cause (past)
effect (future)
ABC
E
A
OR
AND
XOR
B
C
D
composition
blue book
blue
left
book
consciousness is structured: each experience is composed of phenomenological distinctions, elementary or higher-order, which exist within it
the system must be structured: subsets of system elements (composed in various combinations) must have cause–effect power upon the system
[ABC]
higher order mechanism
ON
OFF
[AB]
[AC]
[BC]
[A]
[B]
[C]
elementary mechanisms
information
consciousness is specific: each experience is the particular way it is (it is composed of a specific set of specific phenomenological distinctions), thereby differing from other possible experiences (differentiation)
the system must specify a cause–effect structure that is the particular way it is: a specific set of specific cause–effect repertoires—thereby differing in its specific way from other possible structures (differentiation). A cause–effect repertoire specifies the probability of all possible causes and effects of a mechanism in a state. A cause–effect structure is the set of cause–effect repertoires specified by all subsets of system elements and expresses how the system gives an actual form to the space of possibilities
cause-effect structure
cause repertoires
effect repertoires
integration
consciousness is unified: each experience is irreducible to non-interdependent subsets of phenomenal distinctions
the cause–effect structure specified by the system must be unified: it must be intrinsically irreducible to that specified by non-interdependent sub-systems ($\Phi>0$) across its weakest (unidirectional) link:
MIP = minimum information partition
partitioned cause–effect structure
MIP
exclusion
consciousness is definite, in content and spatio-temporal grain: each experience has the set of phenomenal distinctions it has, not less or more, and flows at the speed it does, not faster or slower
the cause–effect structure specified by the system must be definite: specified over a single set of elements—not lesss or more—and spatio-temporal grains—not faster or slower; this is a cause-effect structure that is maximally irreducible intrinsically ($\Phi^{\max}$), called conceptual structure, made of maximally irreducible cause–effect repertoires (concepts)
$\Phi^{\max}$
time
conceptual structure
cause repertoires
effect repertoires
$\varphi^{\max}$ of concepts
complex
$(\Phi^{\max})=1.92$

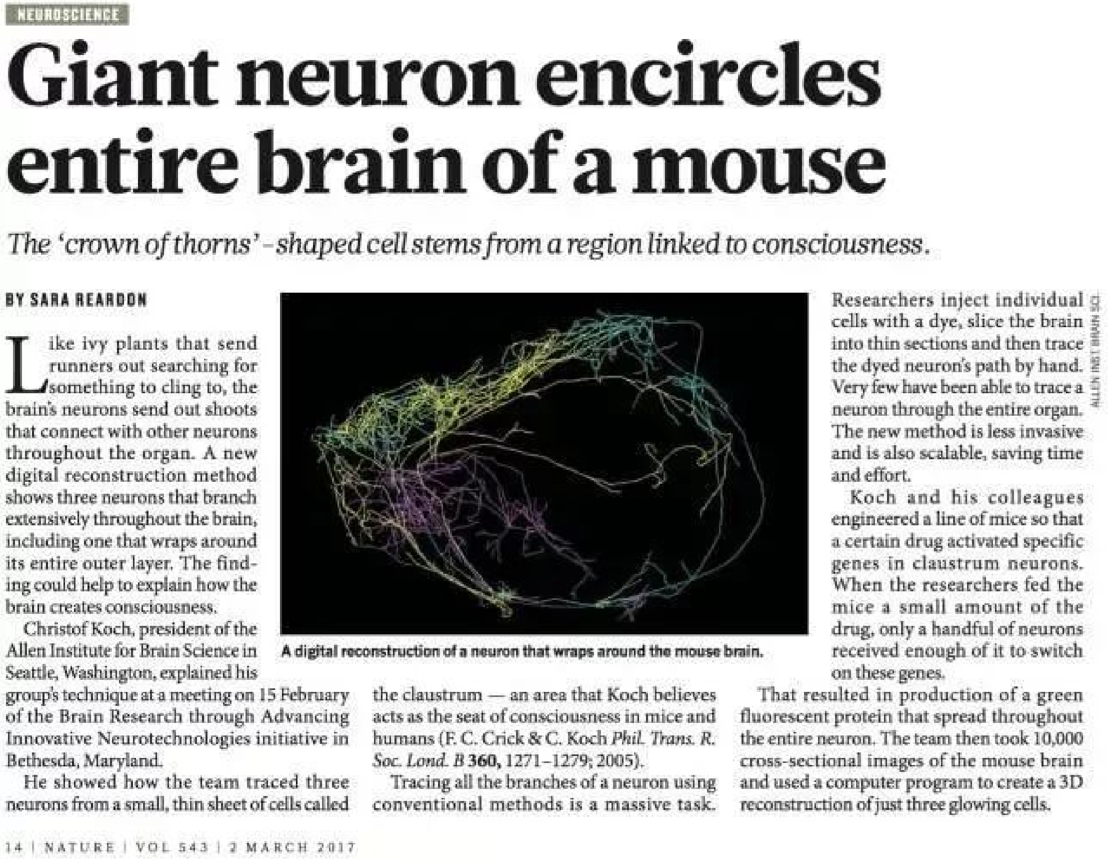
抽象化的表达有自己的优势，因为从量子物理而言，一切都是一个状态，不需要内部模块。

例如 http://xxx.itp.ac.cn/pdf/1712.01826.pdf。将观测者简化为计算过程后，建立了一种既唯心又唯物的宇宙理论。



关于意识的生物基础，近年的著名发现是连接大脑各个部分的巨型神经元。这种神经元，从大脑的屏状核出发，连接到大脑的各个部分，可能是意识的开关。

在一例癫痫患者的人体实验中，确实可通过在屏状核进行高频电刺激，关闭和开启实验者的意识（实验者的感觉就像断片）。



NEUROSCIENCE

Giant neuron encircles entire brain of a mouse

The 'crown of thorns'-shaped cell stems from a region linked to consciousness.

BY SARA REARDON

A digital reconstruction of a neuron that wraps around the mouse brain.

ALLEN INST. BRAIN SCI.

Like ivy plants that send runners out searching for something to cling to, the brain's neurons send out shoots that connect with other neurons throughout the organ. A new digital reconstruction method shows three neurons that branch extensively throughout the brain, including one that wraps around its entire outer layer. The finding could help to explain how the brain creates consciousness.

Christof Koch, president of the Allen Institute for Brain Science in Seattle, Washington, explained his group's technique at a meeting on 15 February of the Brain Research through Advancing Innovative Neurotechnologies initiative in Bethesda, Maryland.

He showed how the team traced three neurons from a small, thin sheet of cells called the claustrum — an area that Koch believes acts as the seat of consciousness in mice and humans (F. C. Crick & C. Koch *Phil. Trans. R. Soc. Lond. B* **360,** 1271–1279; 2005).

Tracing all the branches of a neuron using conventional methods is a massive task. Researchers inject individual cells with a dye, slice the brain into thin sections and then trace the dyed neuron's path by hand. Very few have been able to trace a neuron through the entire organ. The new method is less invasive and is also scalable, saving time and effort.

Koch and his colleagues engineered a line of mice so that a certain drug activated specific genes in claustrum neurons. When the researchers fed the mice a small amount of the drug, only a handful of neurons received enough of it to switch on these genes.

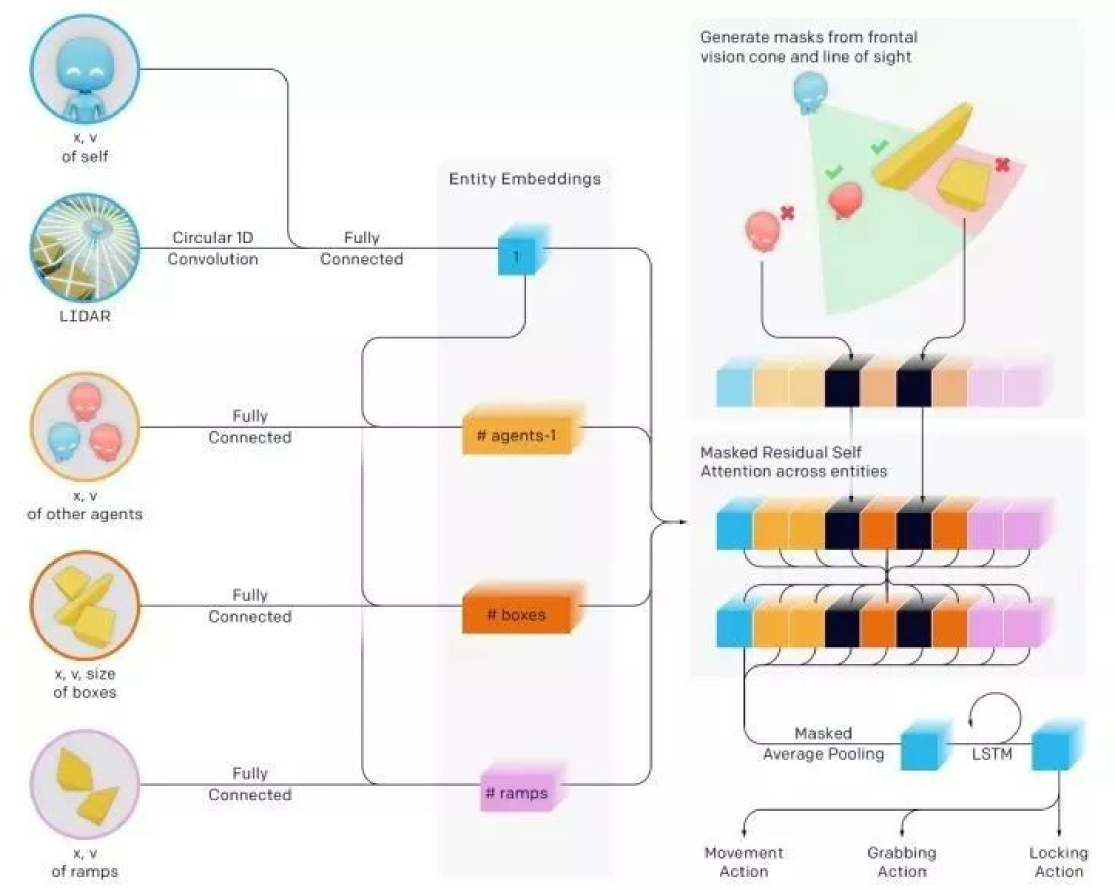
That resulted in production of a green fluorescent protein that spread throughout the entire neuron. The team then took 10,000 cross-sectional images of the mouse brain and used a computer program to create a 3D reconstruction of just three glowing cells.

14 | NATURE | VOL 543 | 2 MARCH 2017

二、在 DRL 实验环境中的模型

为实现良好的多 agent 性能，和产生有趣的合作/竞争行为，目前的 DRL 模型仍需做大量简化，例如采用全局的优化策略。

OpenAI 的 2019 年模型（http://xxx.itp.ac.cn/pdf/1909.07528.pdf）：



OpenAI Five：

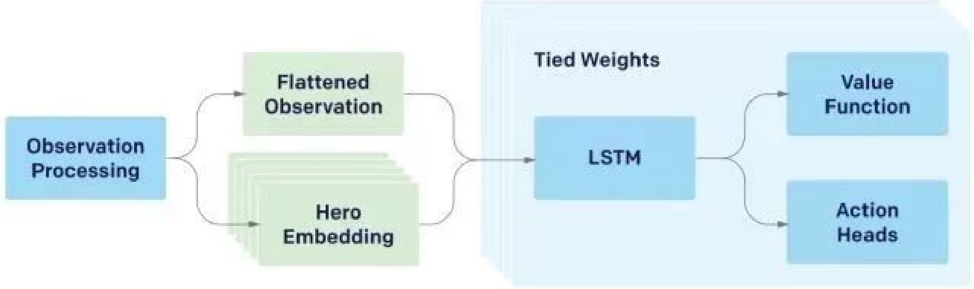
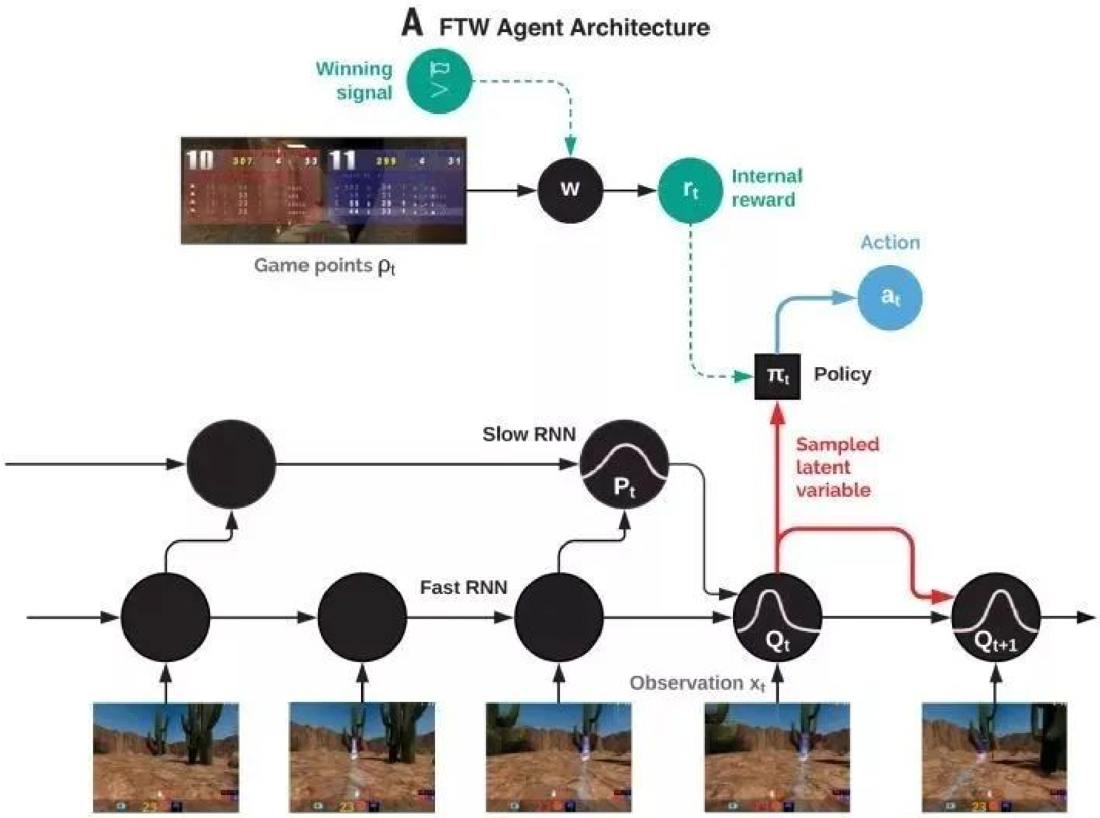
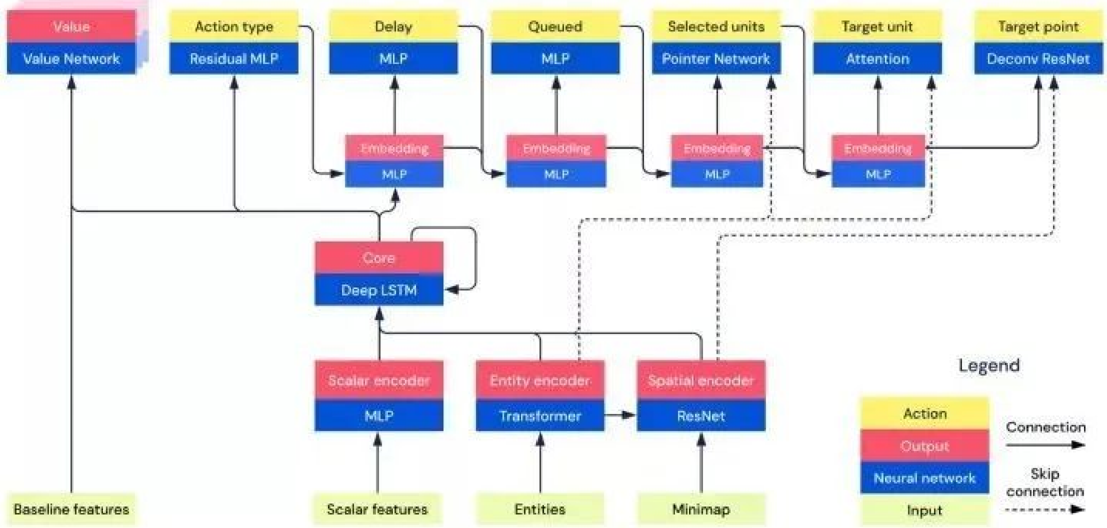


Figure 1: **Simplified OpenAI Five Model Architecture:** The complex multi-array observation space is processed into a single vector, which is then passed through a 4096-unit LSTM. The LSTM state is projected to obtain the policy outputs (actions and value function). Each of the five heroes on the team is controlled by a replica of this network with nearly identical inputs, each with its own hidden state. The networks take different actions due to a part of the observation processing's output indicating which of the five heroes is being controlled. The LSTM composes 84% of the model's total parameter count. See Figure 17 and Figure 18 in Appendix H for a detailed breakdown of our model architecture.

DeepMind 的 2019 年模型（Human-level performance in 3D multiplayer games with populationbased reinforcement learning）：



AlphaStar:



最近流行的论文 A distributional code for value in dopaminebased reinforcement learning ，其中人造的多巴胺神经元可以预测回报的分布。

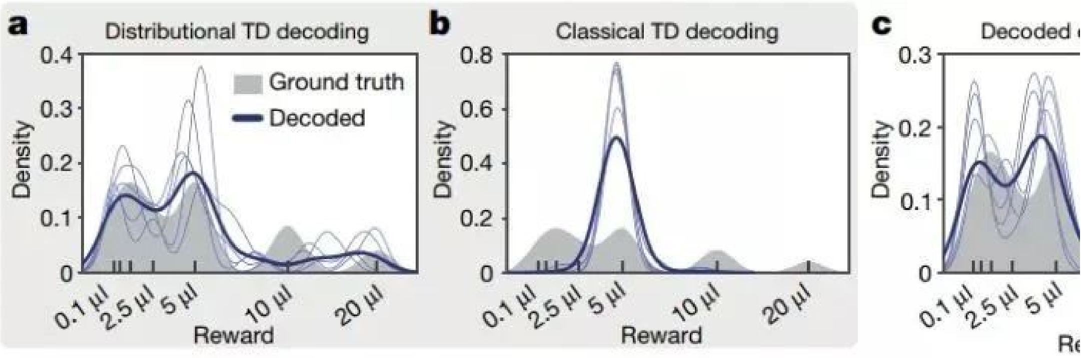
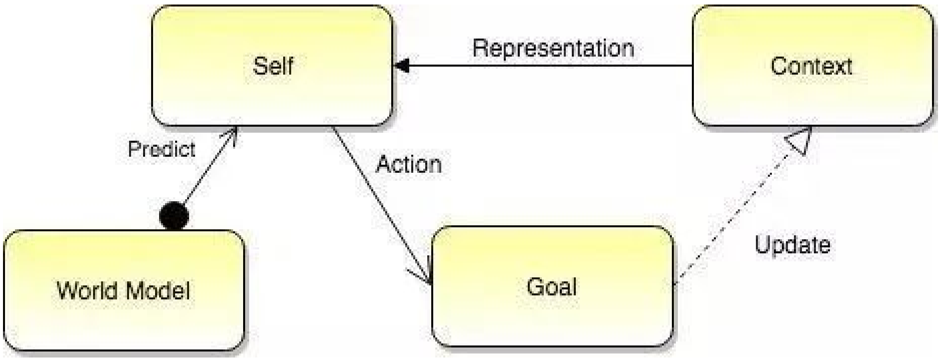
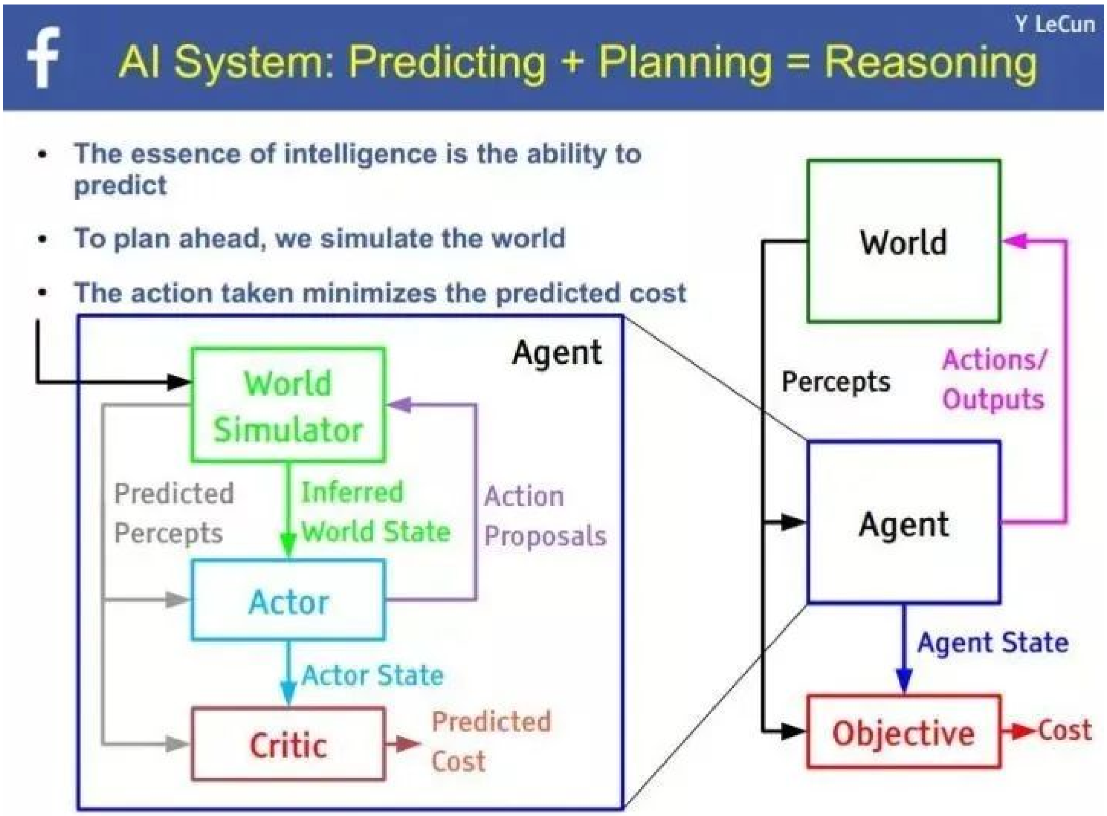


Fig. 5 | Decoding reward distributions from neural responses.
a, Distributional TD simulation trained on the variable-magnitude task, whose actual (smoothed) distribution of rewards is shown in grey. After training the model, we interpret the learned values as a set of expectiles. We then decode the set of expectiles into a probability density (blue traces). Multiple solutions are shown in light blue, and the average across solutions is shown in dark blue.

三、世界模型

从更广泛的观点看，大脑会建立世界模型：

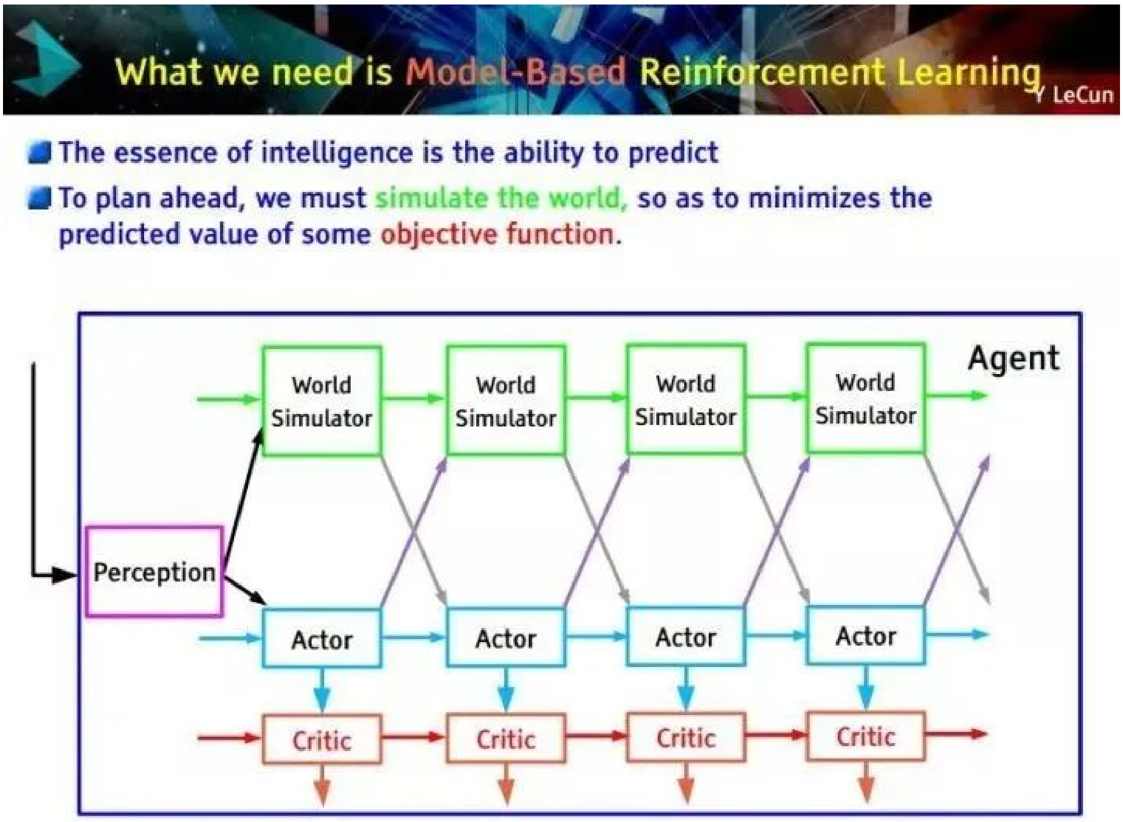




Y LeCun

AI System: Predicting + Planning = Reasoning

- The essence of intelligence is the ability to predict
- To plan ahead, we simulate the world
- The action taken minimizes the predicted cost

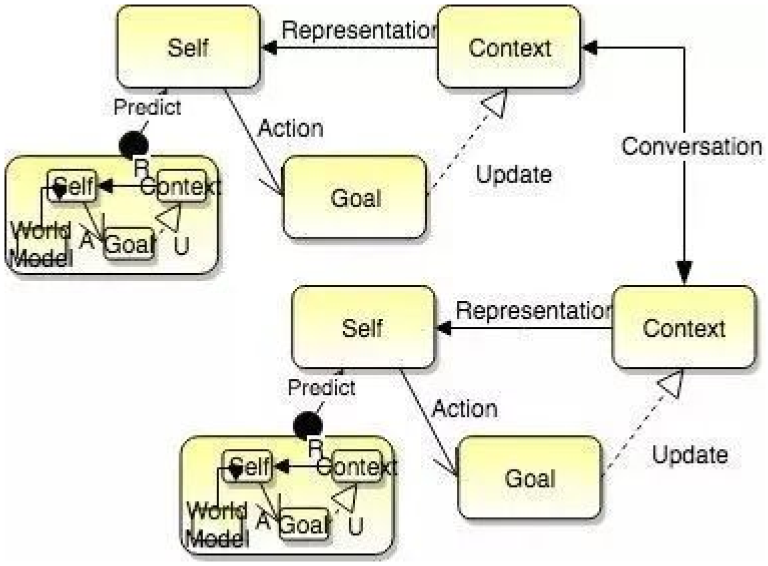


What we need is Model-Based Reinforcement Learning

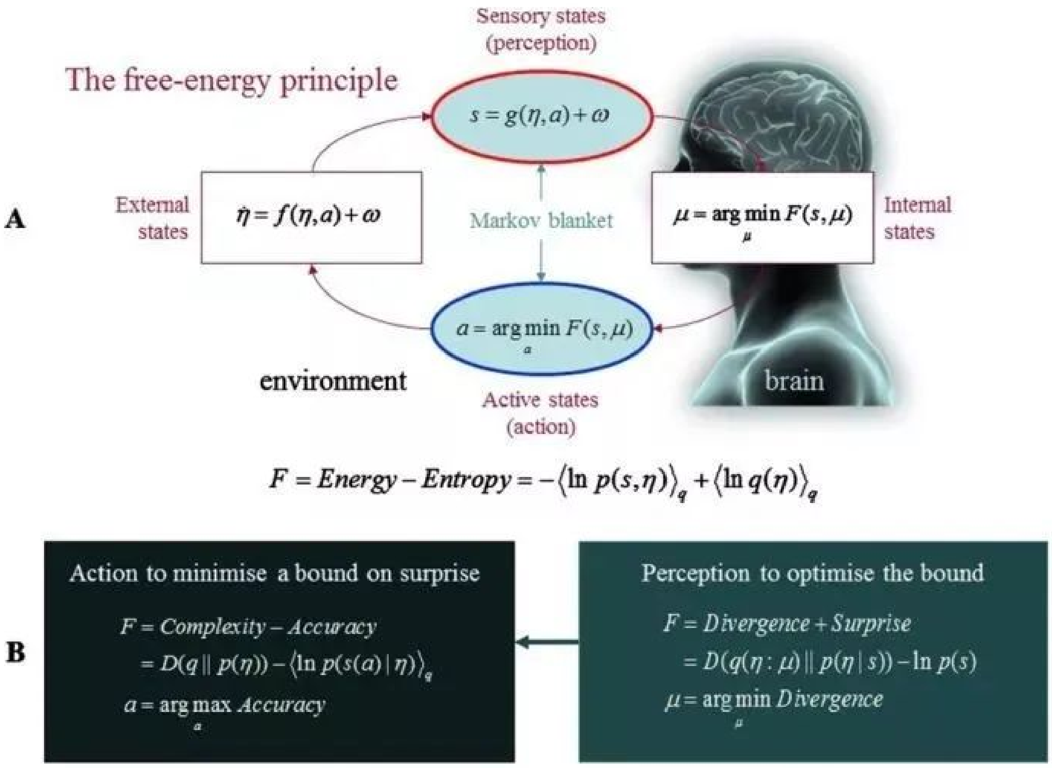
Y LeCun

- The essence of intelligence is the ability to predict
- To plan ahead, we must simulate the world, so as to minimizes the predicted value of some objective function.

多 agent 可进行交流：



从“心智中的世界模型”出发，也可建立现实的理论。例如 Karl Friston 提出的自由能理论，近年引人注目：

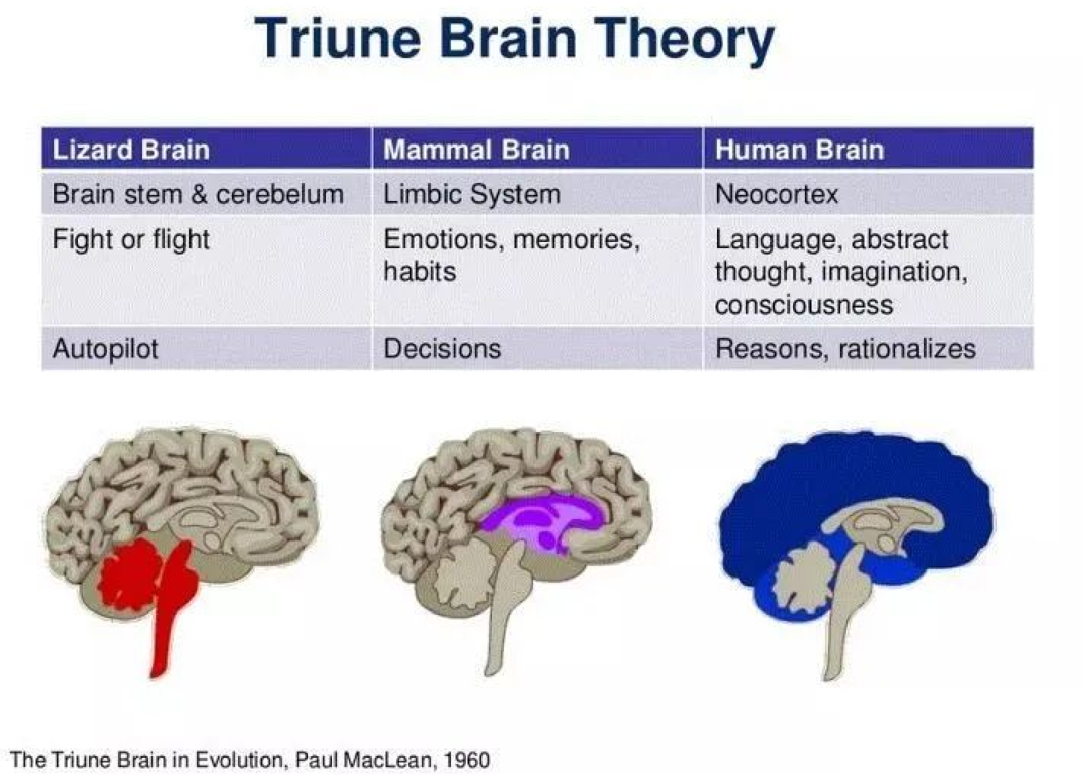

Sensory states
(perception)
The free-energy principle
$s = g(\eta, a) + \omega$
A
External states
$\dot{\eta} = f(\eta, a) + \omega$
Markov blanket
$\mu = \arg\min_{\mu} F(s, \mu)$
Internal states
$a = \arg\min_{a} F(s, \mu)$
environment
Active states
(action)
brain
$F = Energy - Entropy = -\langle \ln p(s, \eta) \rangle_q + \langle \ln q(\eta) \rangle_q$
B
Action to minimise a bound on surprise
$F = Complexity - Accuracy$
$= D(q \| p(\eta)) - \langle \ln p(s(a) | \eta) \rangle_q$
$a = \arg\max_{a} Accuracy$
Perception to optimise the bound
$F = Divergence + Surprise$
$= D(q(\eta : \mu) \| p(\eta | s)) - \ln p(s)$
$\mu = \arg\min_{\mu} Divergence$

它认为，生物的本质在于让世界模型吻合世界：一面修改世界模型以符合世界（这是显然的），一面修改世界以符合世界模型（这是有趣且也有道理的）。总而言之，降低熵。

四、心理与非理性模型，精神分析

上文关注机械的理性与决策，只代表大脑的一部分。非理性和无意识的部分，同样值得考虑。

生物学上有 Triune Brain 理论，将大脑分为 生存大脑 - 情绪大脑 - 理性大脑（实际情况比这更复杂）：

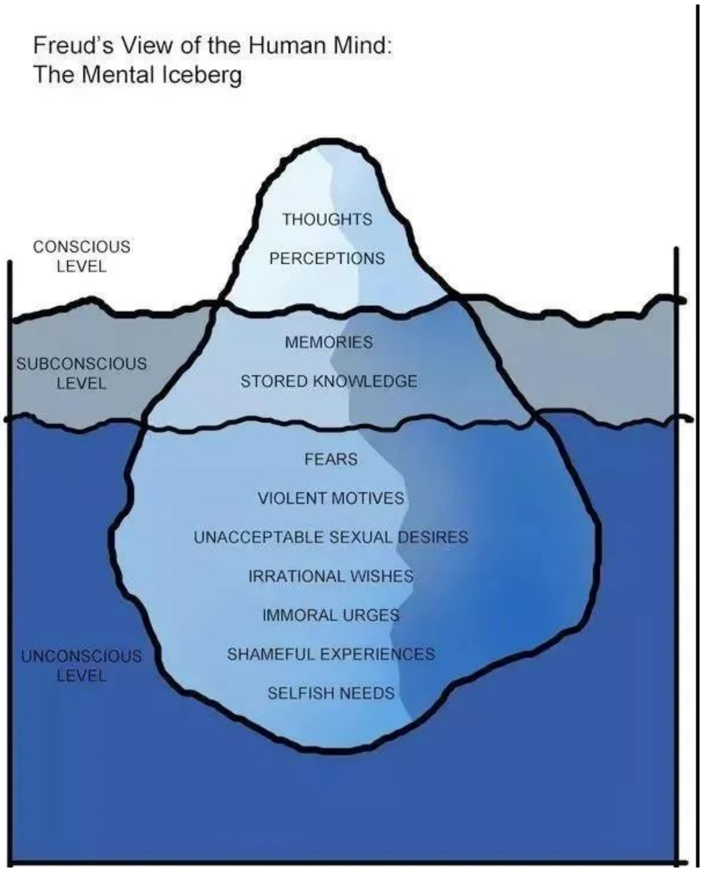


Triune Brain Theory

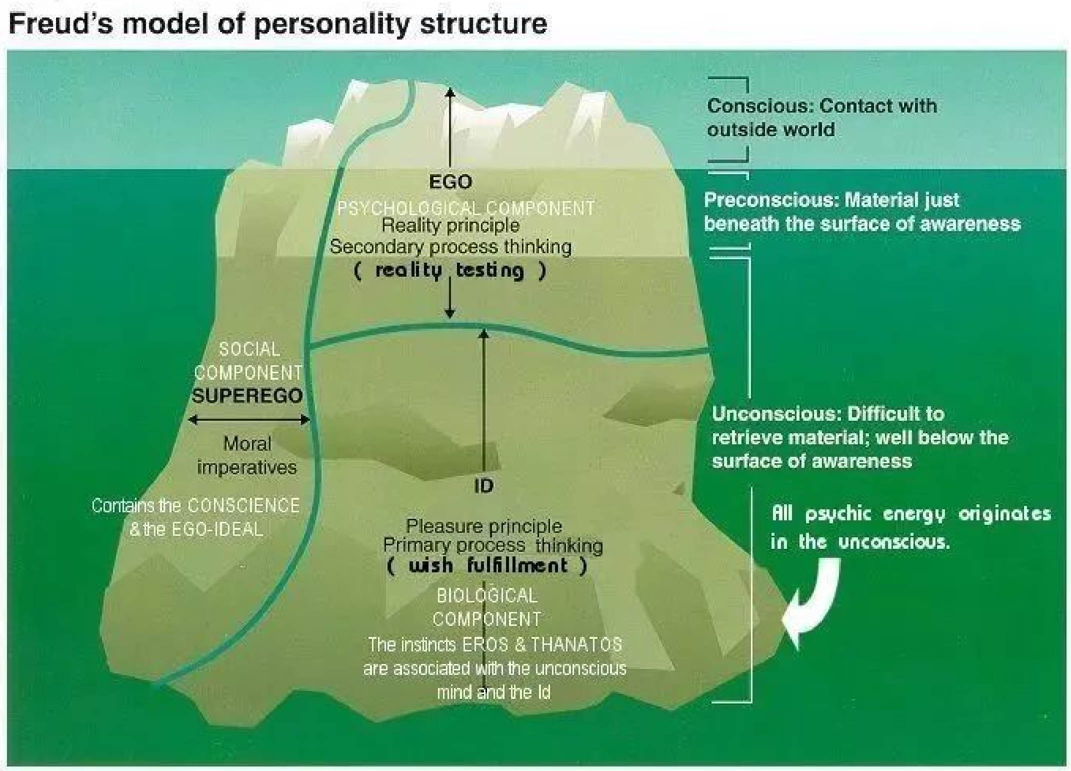
Lizard Brain	Mammal Brain	Human Brain
Brain stem & cerebelum	Limbic System	Neocortex
Fight or flight	Emotions, memories, habits	Language, abstract thought, imagination, consciousness
Autopilot	Decisions	Reasons, rationalizes

The Triune Brain in Evolution, Paul MacLean, 1960

人的思维是从无意识而来。弗洛伊德的第一拓比（无意识 - 潜意识 - 意识）：



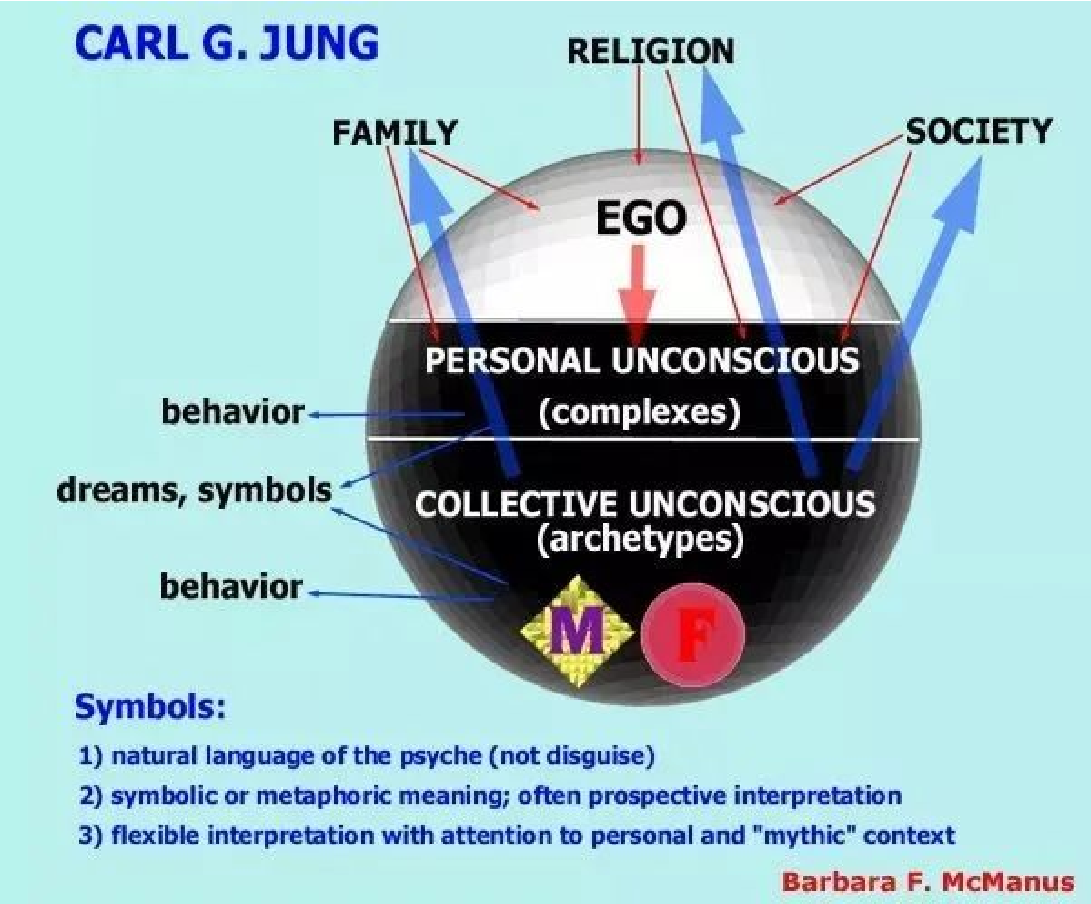
弗洛伊德的第二拓比（本我 – 自我 – 超我），这里开始体现社会化/符号界：


Freud's model of personality structure
Conscious: Contact with outside world
Preconscious: Material just beneath the surface of awareness
Unconscious: Difficult to retrieve material; well below the surface of awareness
All psychic energy originates in the unconscious.
EGO
PSYCHOLOGICAL COMPONENT
Reality principle
Secondary process thinking
(reality testing)
SOCIAL COMPONENT
SUPEREGO
Moral imperatives
Contains the CONSCIENCE & the EGO-IDEAL
ID
Pleasure principle
Primary process thinking
(wish fulfillment)
BIOLOGICAL COMPONENT
The instincts EROS & THANATOS are associated with the unconscious mind and the Id

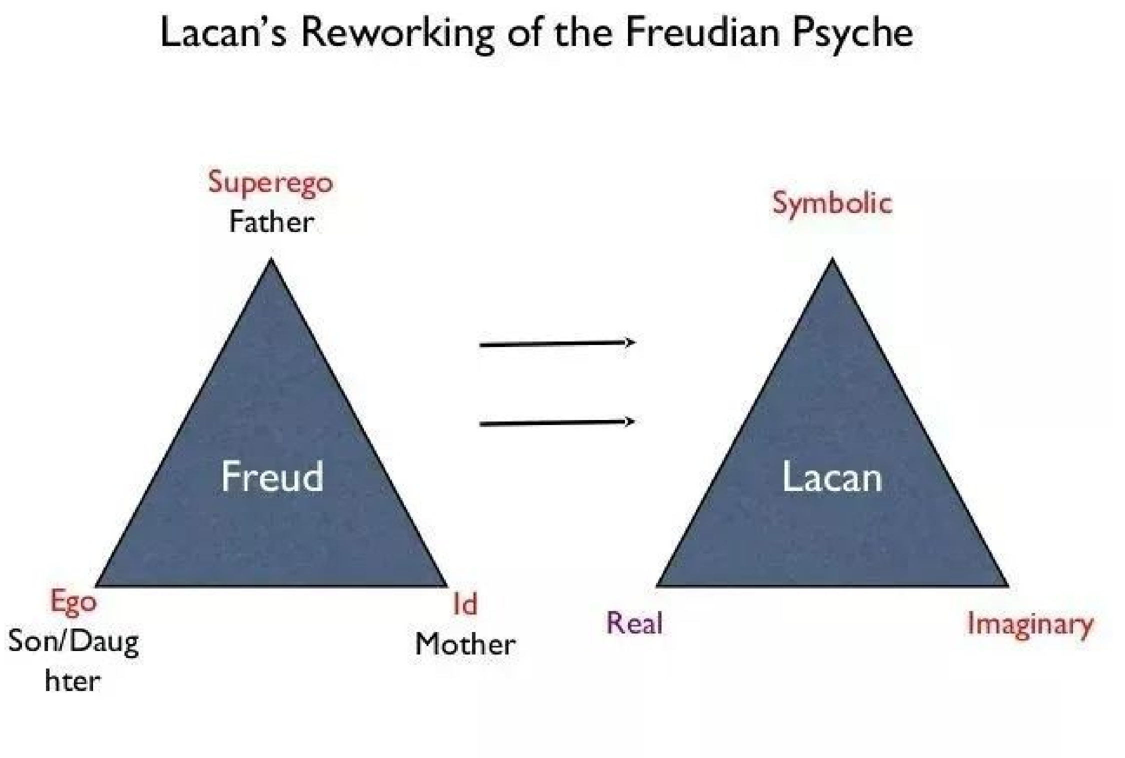
这类视角有意义。因为前文的模型，更像描述动物的行为：动物也会合作和竞争，但都很原始。

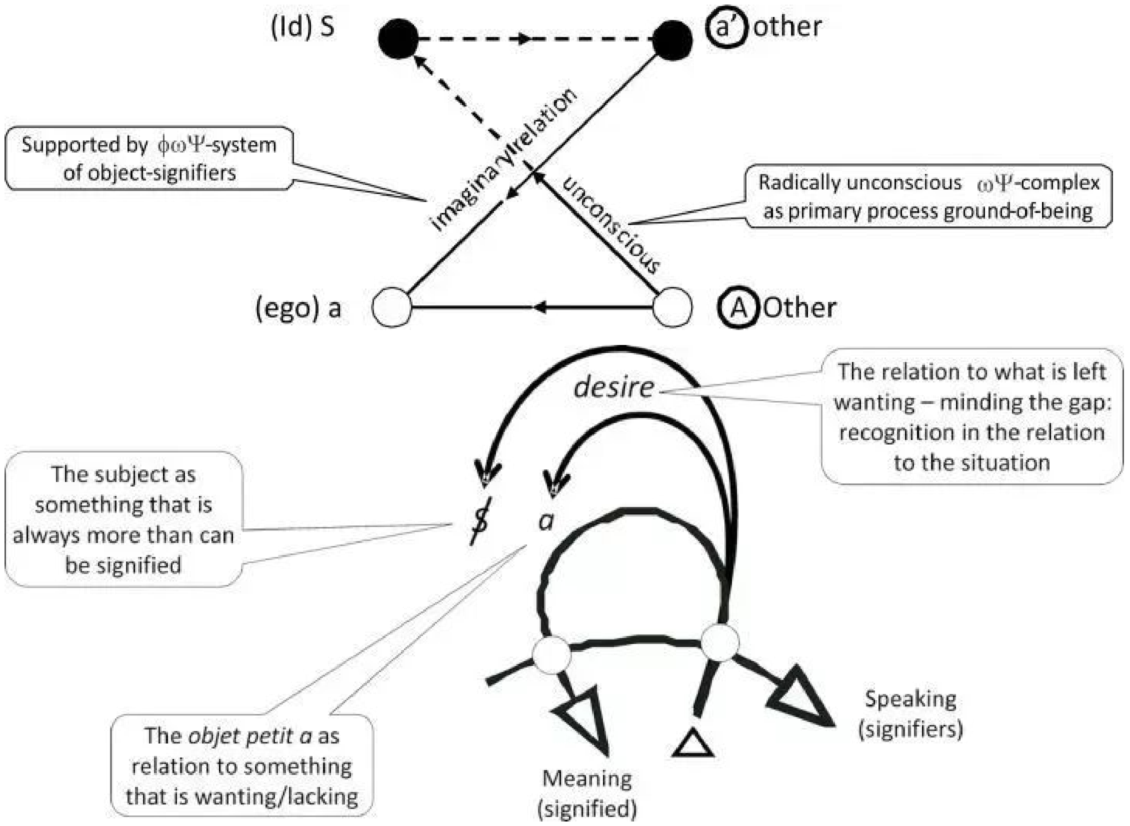
而人类已经高度社会化。语言/文化/社会结构/意识形态/MEME 等等，形成了外部记忆/外部系统/外部意识，并深刻塑造人类的行为。就像那个著名的笑话：人类不但是基因繁衍的方法，人类还是汽车繁衍的方法。如果从还原论看，可认为这些仍然可来自此前的模型，不过还原论是否正确，现在不知道，所以还是结合多种观点更为有趣。

荣格（意识 - 个人无意识 - 集体无意识），这里的无意识仍然是神秘的混沌冲动：

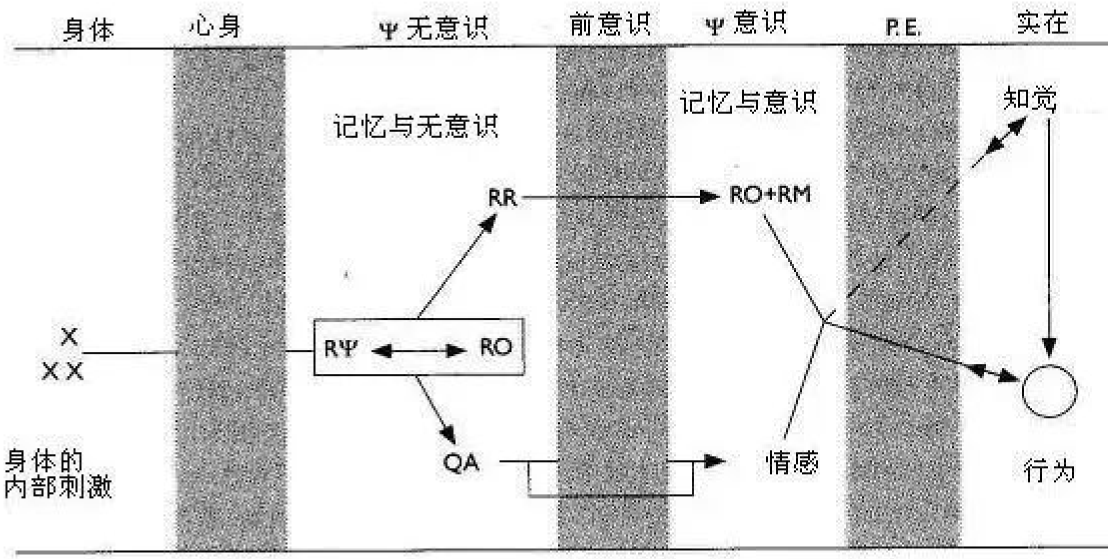

CARL G. JUNG
RELIGION
FAMILY
SOCIETY
EGO
PERSONAL UNCONSCIOUS
(complexes)
behavior
dreams, symbols
COLLECTIVE UNCONSCIOUS
(archetypes)
behavior
M
F
Symbols:
1) natural language of the psyche (not disguise)
2) symbolic or metaphoric meaning; often prospective interpretation
3) flexible interpretation with attention to personal and "mythic" context
Barbara F. McManus

拉康（实在界 - 想象界 - 符号界），这里的无意识是更清晰的，来自于他者：


Lacan's Reworking of the Freudian Psyche


(Id) S
a'
other
Supported by ϕωΨ-system of object-signifiers
imaginary relation
unconscious
Radically unconscious ωΨ-complex as primary process ground-of-being
(ego) a
A
Other
desire
The relation to what is left wanting – minding the gap: recognition in the relation to the situation
The subject as something that is always more than can be signified
$
a
The objet petit a as relation to something that is wanting/lacking
Speaking (signifiers)
Meaning (signified)

后续的客体关系学派的例子：



RΨ = 冲动的精神代表

RR = 代表-表象

QA = 情感当量

RO = 物表象或客体表象
（意识或无意识）

RM = 词表象

O = 客体

灰色区域分别代表：

心身边界
（在身体和无意识之间）

前意识的阻障
（代表与之同在）

P.E.=泛兴奋的区域

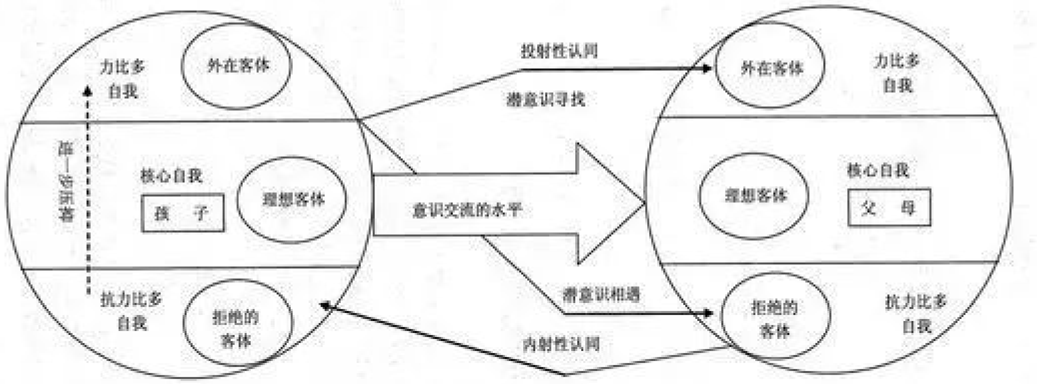
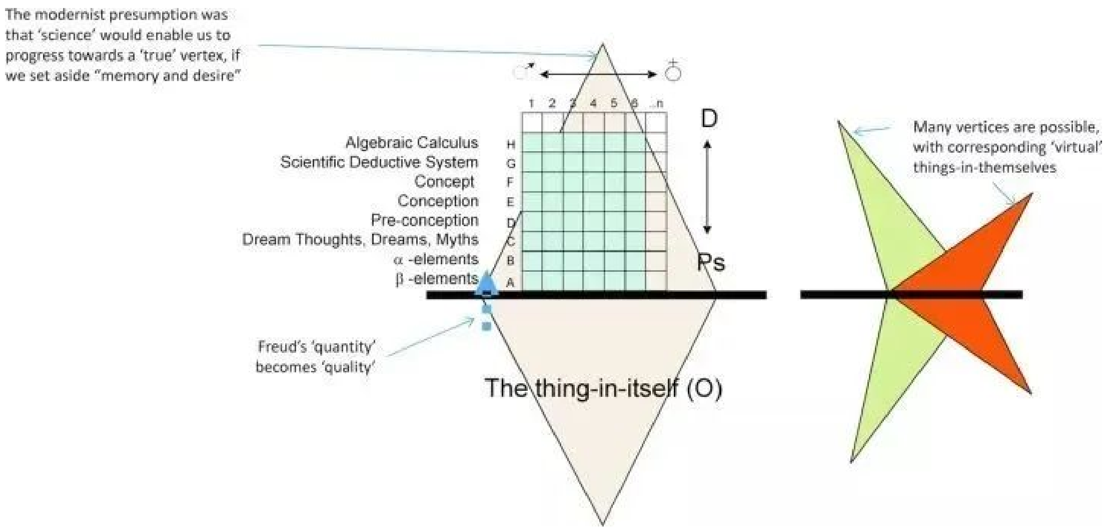


图5.1　母婴关系中的投射性和内射性认同。这里的机制是当婴儿遇到沮丧的、无回报的渴望，或创伤时，孩子与父母的投射性和内射性认同之间的交流。该图描述了孩子渴望的需要被满足，通过投射性认同与父母相似的趋向认同。遭遇到拒绝的孩子便通过内射性认同与父母内心中抗力比多系统的沮丧进行了认同。在对沮丧的内在反应中，力比多系统受到孩子的抗力比多系统的力量的进一步压抑。摘自《性关系：性和家庭的客体关系观点》，由Routledge和Kegan Paul共同授权。版权属David E. Scharff，1982。

实际上有很多有趣的想法，例如：

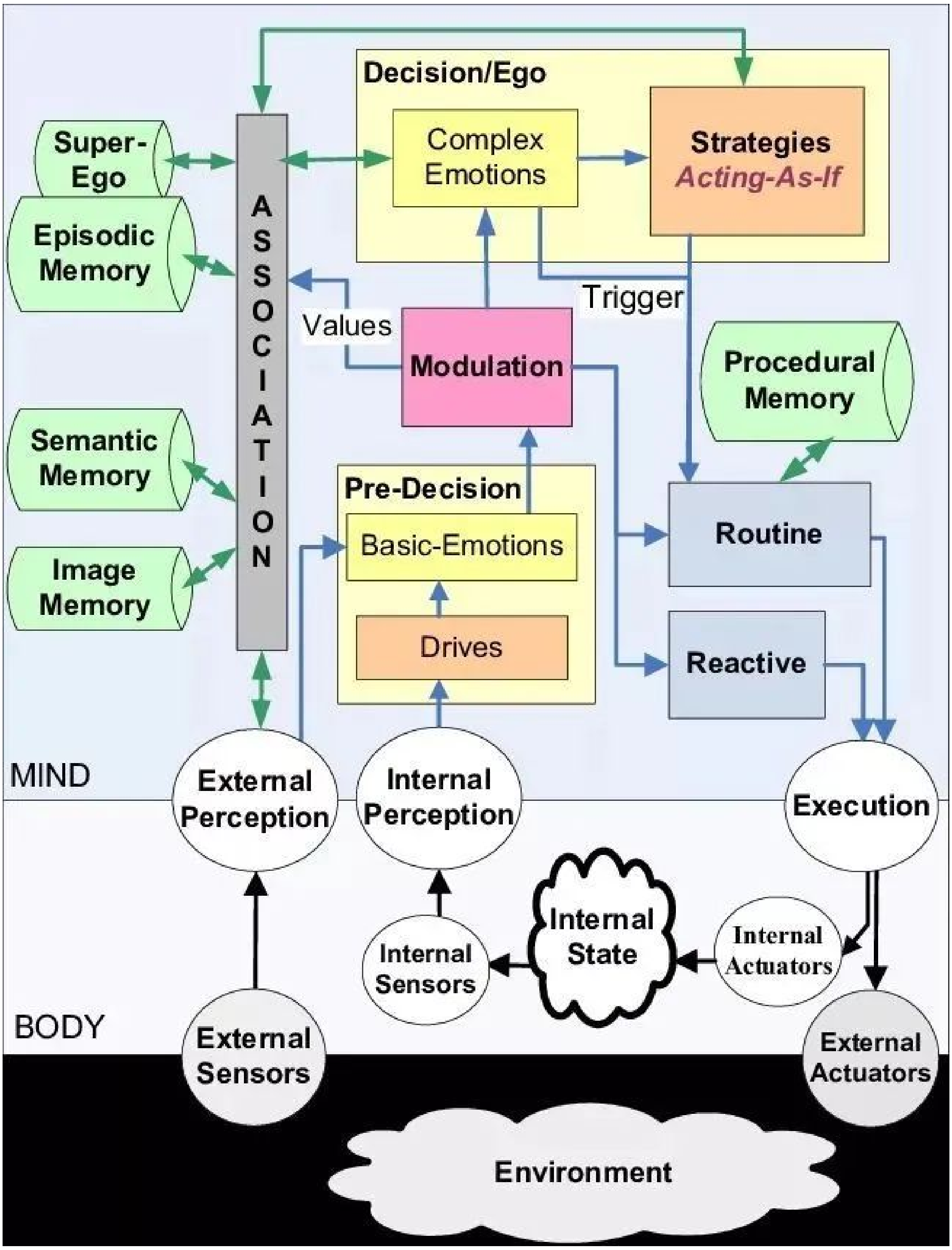

The modernist presumption was that 'science' would enable us to progress towards a 'true' vertex, if we set aside "memory and desire"
♂
♀
1 2 3 4 5 6 ...n
D
Algebraic Calculus H
Scientific Deductive System G
Concept F
Conception E
Pre-conception D
Dream Thoughts, Dreams, Myths C
α -elements B
β -elements A
Ps
Freud's 'quantity' becomes 'quality'
The thing-in-itself (O)
Many vertices are possible, with corresponding 'virtual' things-in-themselves

怎么让 AI 理解这里的这些，是个 NLP 的难题。

五、融合

自然的想法，是将这里的观点，包括量子物理（观测问题仍然是复杂的）和数学和各种哲学理论，全部融合。

例如，一种 1+4 的初步融合例子：


Decision/Ego
Complex Emotions
Strategies
Acting-As-If
Super-Ego
Episodic Memory
ASSOCIATION
Trigger
Values
Modulation
Procedural Memory
Semantic Memory
Pre-Decision
Routine
Basic-Emotions
Image Memory
Drives
Reactive
MIND
External Perception
Internal Perception
Execution
Internal State
Internal Actuators
Internal Sensors
BODY
External Sensors
External Actuators
Environment