Функциональные возможности биологически инспирированных когнитивных архитектур

Алексей В. Самсонович
Университет Джорджа Мэйсона
г. Фэйрфакс, Вирджиния, США

Нейроинформатика – 2015: Москва, МИФИ, 19-23 января

Функциональные возможности биологически инспирированных когнитивных архитектур

- *Аннотация*: Основные функциональные аспекты мышления человека можно описать на вычислительном уровне и воспроизвести в машине на принципах не требующих детального моделирования нейронов и структур мозга. Это прежде всего основные принципы восприятия и осмысления информации, принятия и исполнения решений, сформулированные на языке символьных моделей типа когнитивных архитектур. Ключевыми же являются принципы социально-эмоционального интеллекта, нарративного интеллекта, мета-мышления, автономного выбора целей, семантического картирования, человекоподобной обучаемости и креативности. Создание в машине аналога человеческого субъекта на этих принципах и признание его людьми на уровне равного человеку персонажа приведет к технологическому прорыву, который окажет влияние на все сферы жизни человека.
- Ключевые слова: когнитивные архитектуры; нарративный интеллект; социально-эмоциональное мышление

Четыре научных направления сводящиеся к одной задаче:

1

Вычислительная нейронаука

- *Goal:* Parsimoniously explain in detail **how the brain works,** Reverse-engineer the brain

2

Когнитивное моделирование

- *Goal:* describe computationally, and be able to predict in detail, **human behavior and underlying cognitive processes**

3

Искусственный интеллект

- *Goal:* Develop a **practically useful artificial intelligence** that will replace and outperform humans in a broad spectrum of valuable cognitive tasks

4

Машинное сознание

- *Goal:* Generally “conscious” artifacts capable of becoming **useful members of the human society**

Четыре разрыва в возможностях и подходах

- Разрыв в понимании процессов лежащих в основе мышления на высшем и на элементарном уровне
- Разрыв между искусственным и естественным интеллектом
- Разрыв между железом и ПО
- Разрыв между ролью человека и искусственного интеллектуального агента

Разрыв в вычислительных подходах:

- Logical circuits, control systems (cybernetics)
- Nonlinear dynamical systems, self-organization
- Models of neurons and their parts
- Neuronal networks (connectionism)

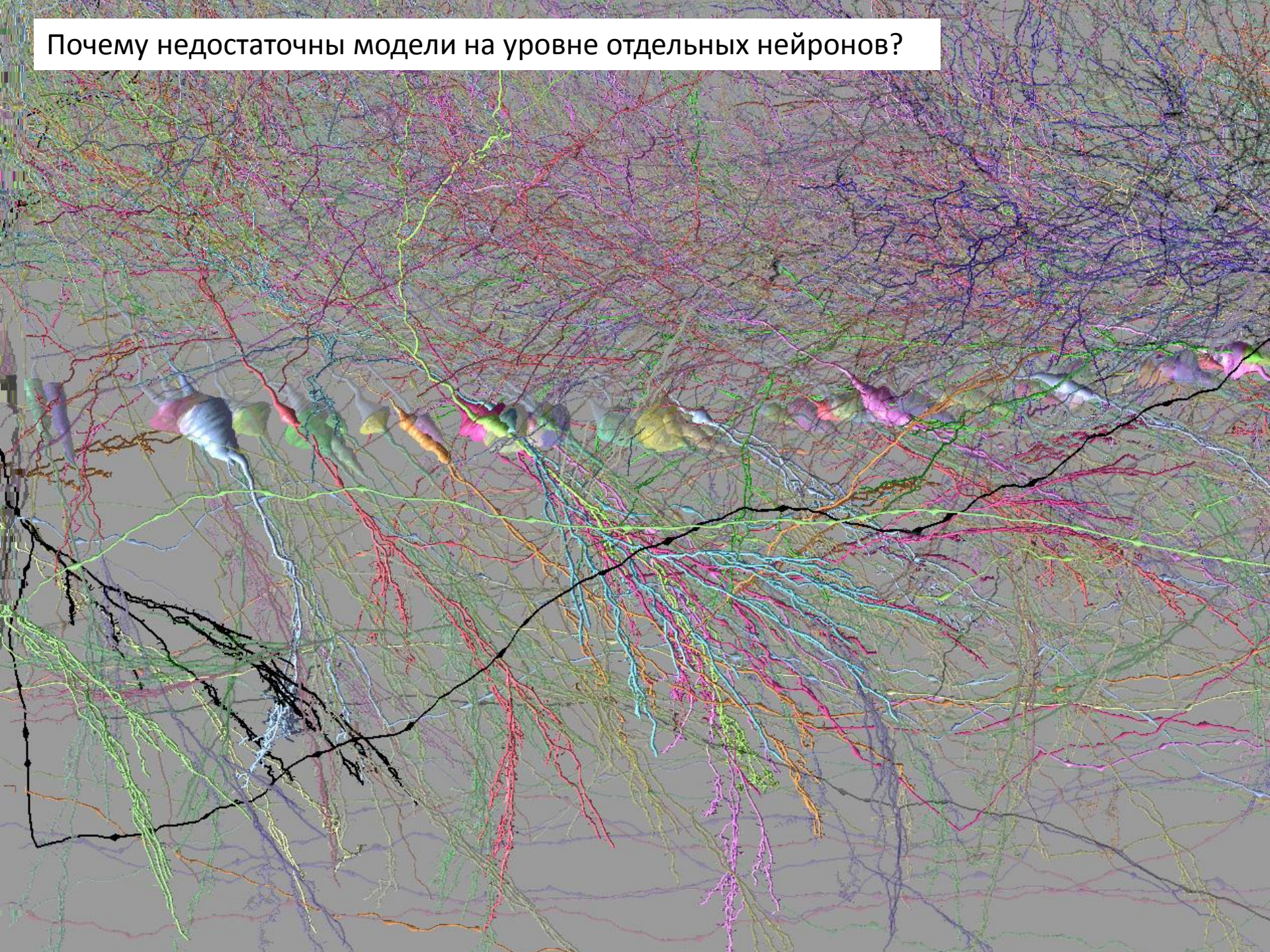
- GAP

- Cognitive neuropsychological models of agency, etc.
- Symbolic modeling
- Modal logic, event calculus, etc.
- Expert systems, automated reasoners & planners
- Intelligent agents, virtual characters, social modeling

Возможно интересный взгляд:

- Для моделирования когнитивных функций человеческого мозга не требуется нейронный уровень
- Когнитивные карты – одна из основ интеллекта
- Биологически инспнрнрованные когнитивные архитектуры (BICA) выступают как объединяющая парадигма требующая комплексного подхода
- Ключевыми являются принципы социально-эмоционального и нарративного интеллекта (включая мета-мышление, автономный выбор целей, …), и человекоподобная обучаемость

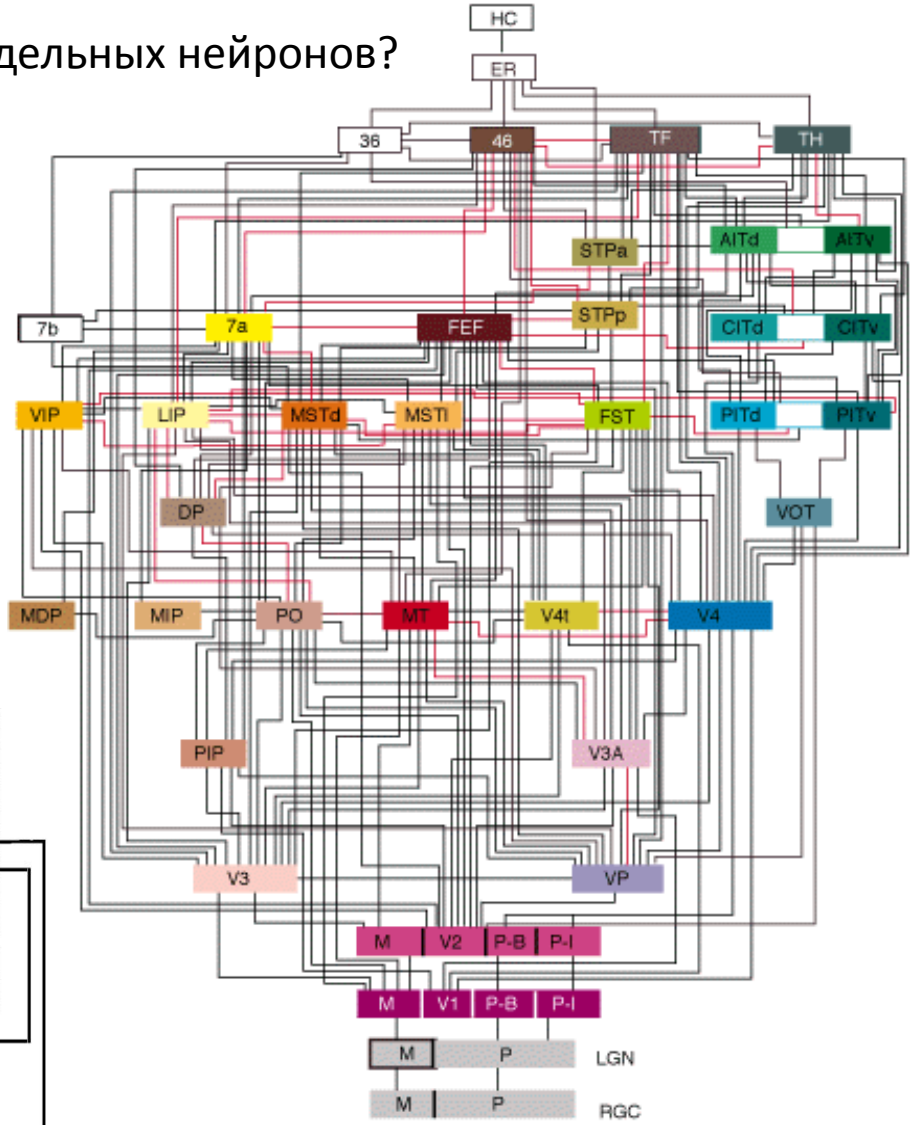
Почему недостаточны модели на уровне отдельных нейронов?



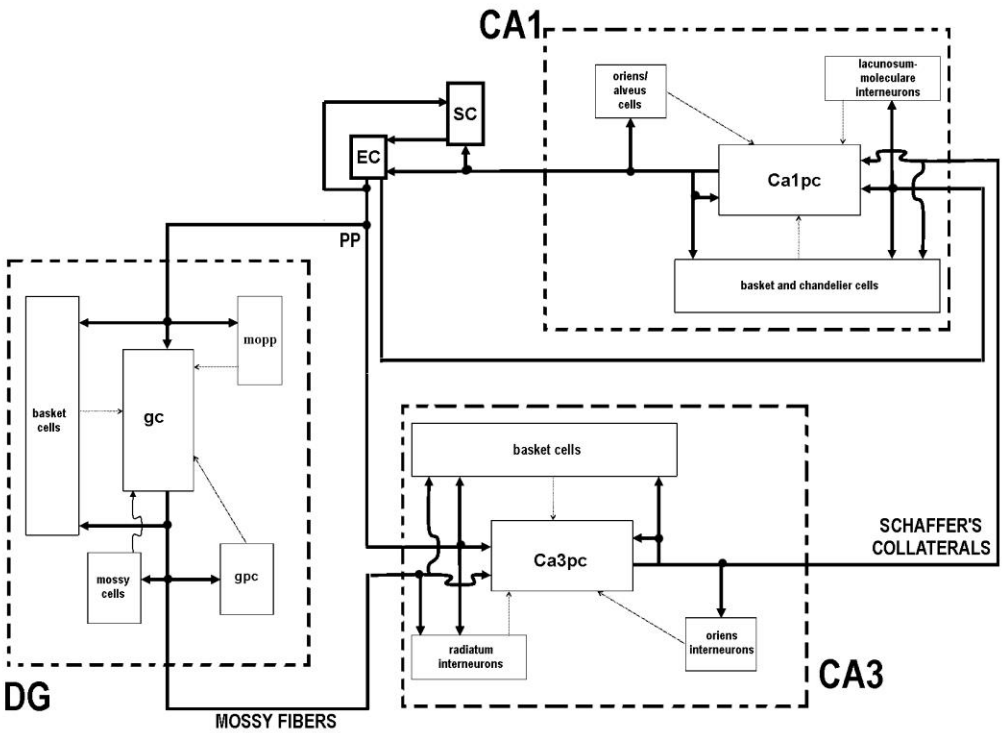
Почему недостаточны модели на уровне отдельных нейронов?

Wire diagram of the visual system

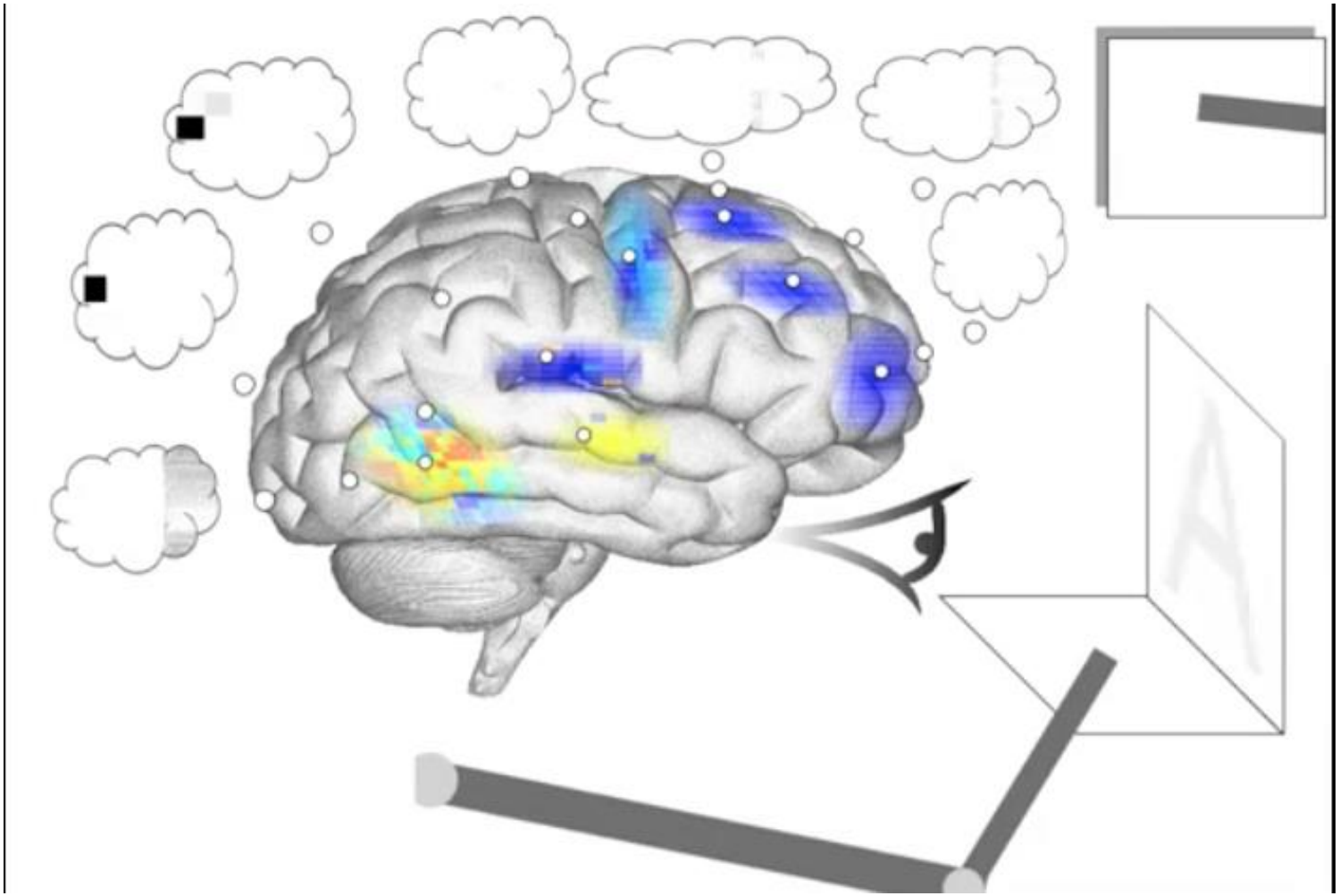
Hierarchy of the visual areas in the brain of a macaque monkey. From: Felleman, D.J. and Van Essen, D.C. (1991)



Hippocampal connectivity diagram



Почему недостаточны модели на уровне отдельных нейронов?



A Large-Scale Model of the Functioning Brain

Chris Eliasmith, Terrence C. Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, Daniel Rasmussen. Science 30 November 2012: Vol. 338 no. 6111 pp. 1202-1205 DOI: 10.1126/science.1225266

Что нужно для преодоления разрыва?

<u>Levels of paradigms in understanding intelligence:</u>

- Psychic and social phenomena

↑

- Cognitive psychological models

- ...

↑

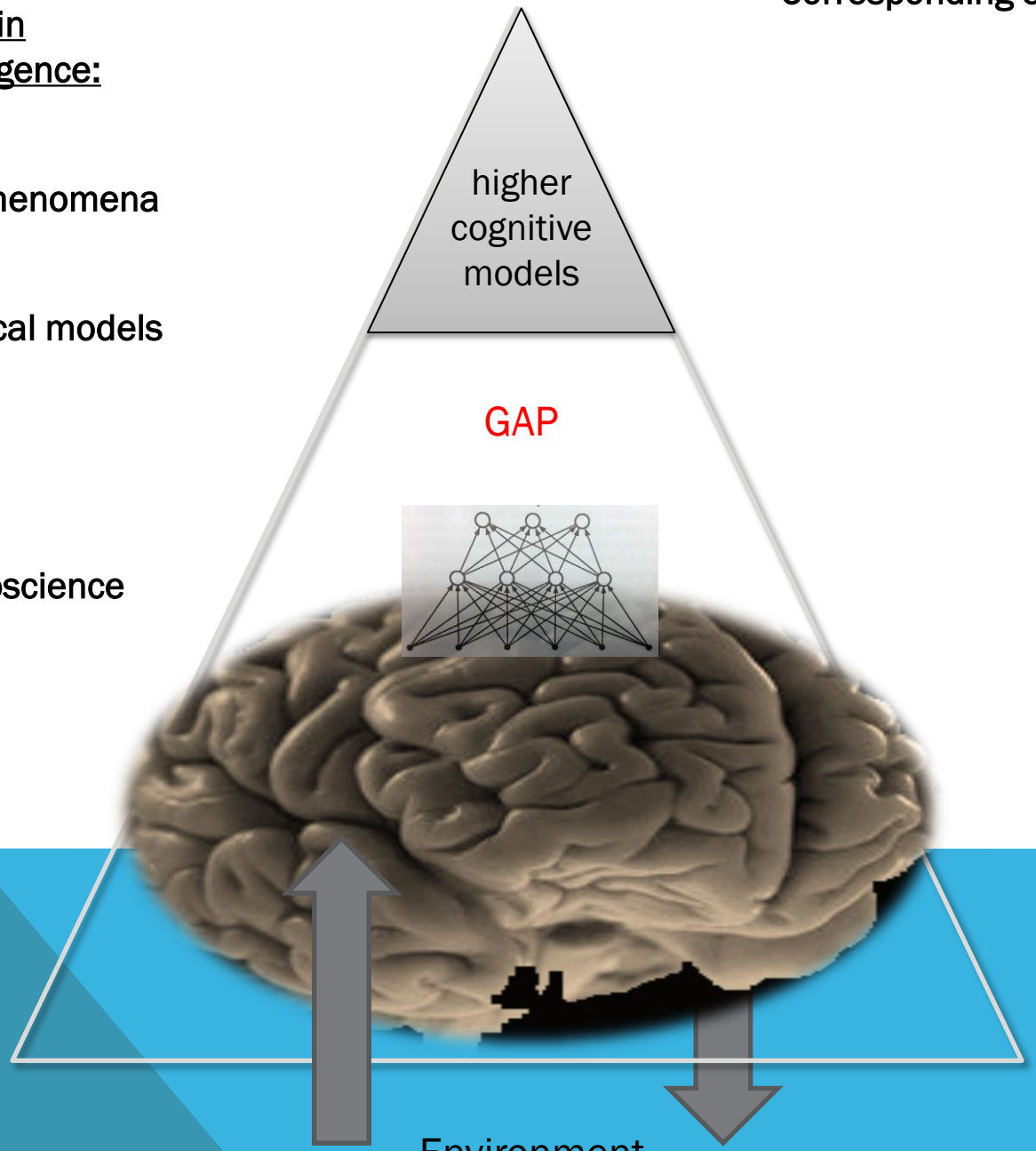
- Computational neuroscience

↑

- ‘Black box’ AI

Corresponding elements:

The Self
Personality
Values...

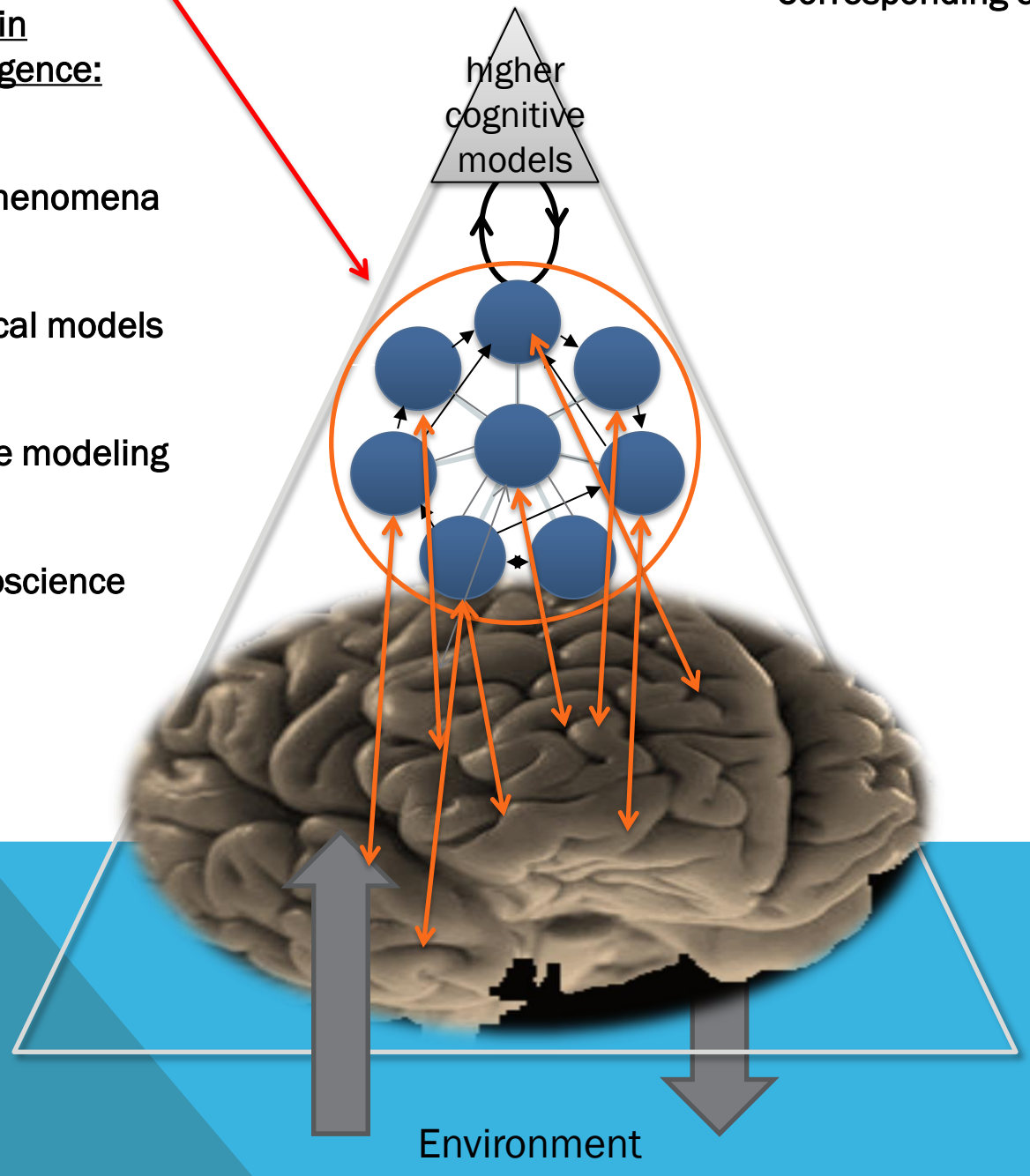


Neurons
Synapses
Neural networks
Patterns...

BICA

<u>Levels of paradigms in understanding intelligence:</u>

- Psychic and social phenomena
- ↑
- Cognitive psychological models
- ↑
- Brain-based cognitive modeling
- ↑
- Computational neuroscience
- ↑
- 'Black box' AI



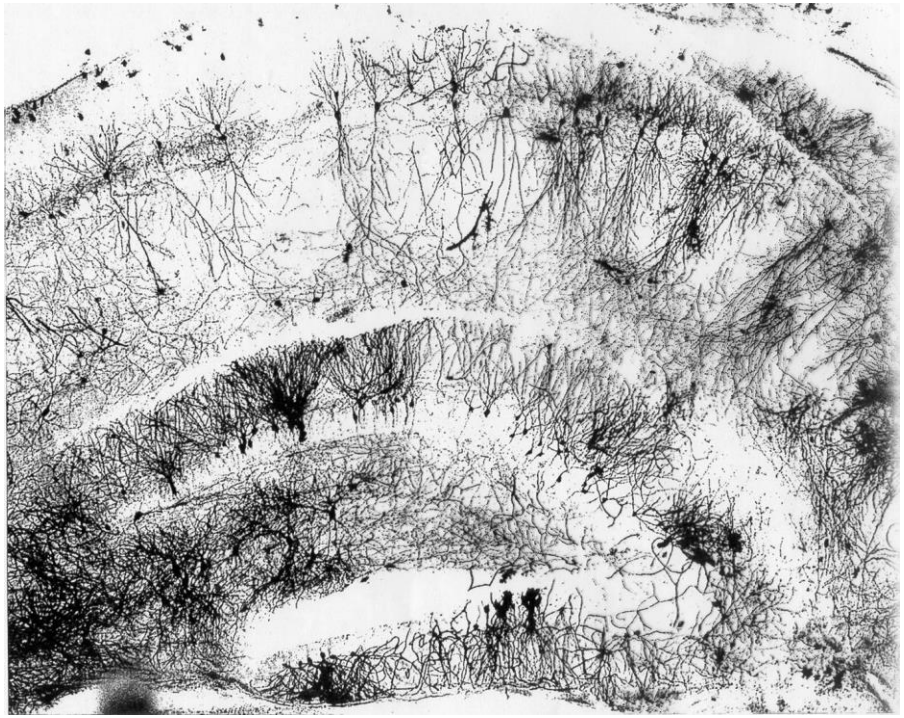
Corresponding elements:

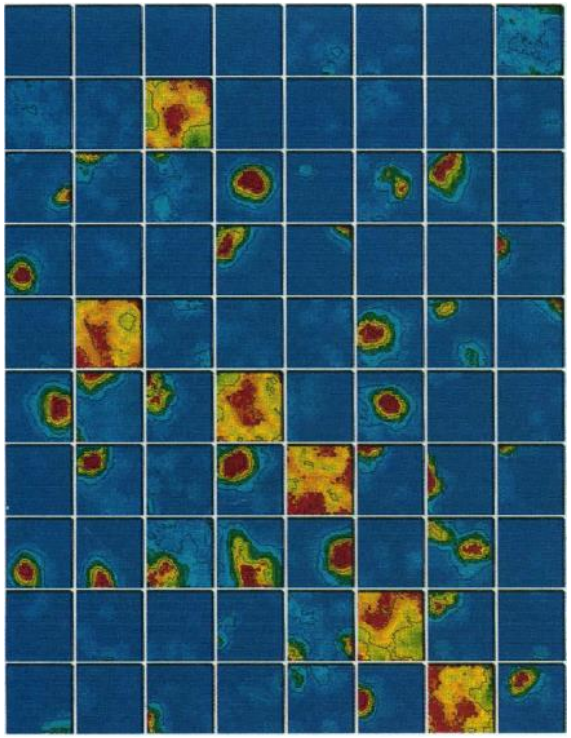
The Self
Personality
Values...

Mental states
Schemas
Frames
Rules
Chunks
Concepts...

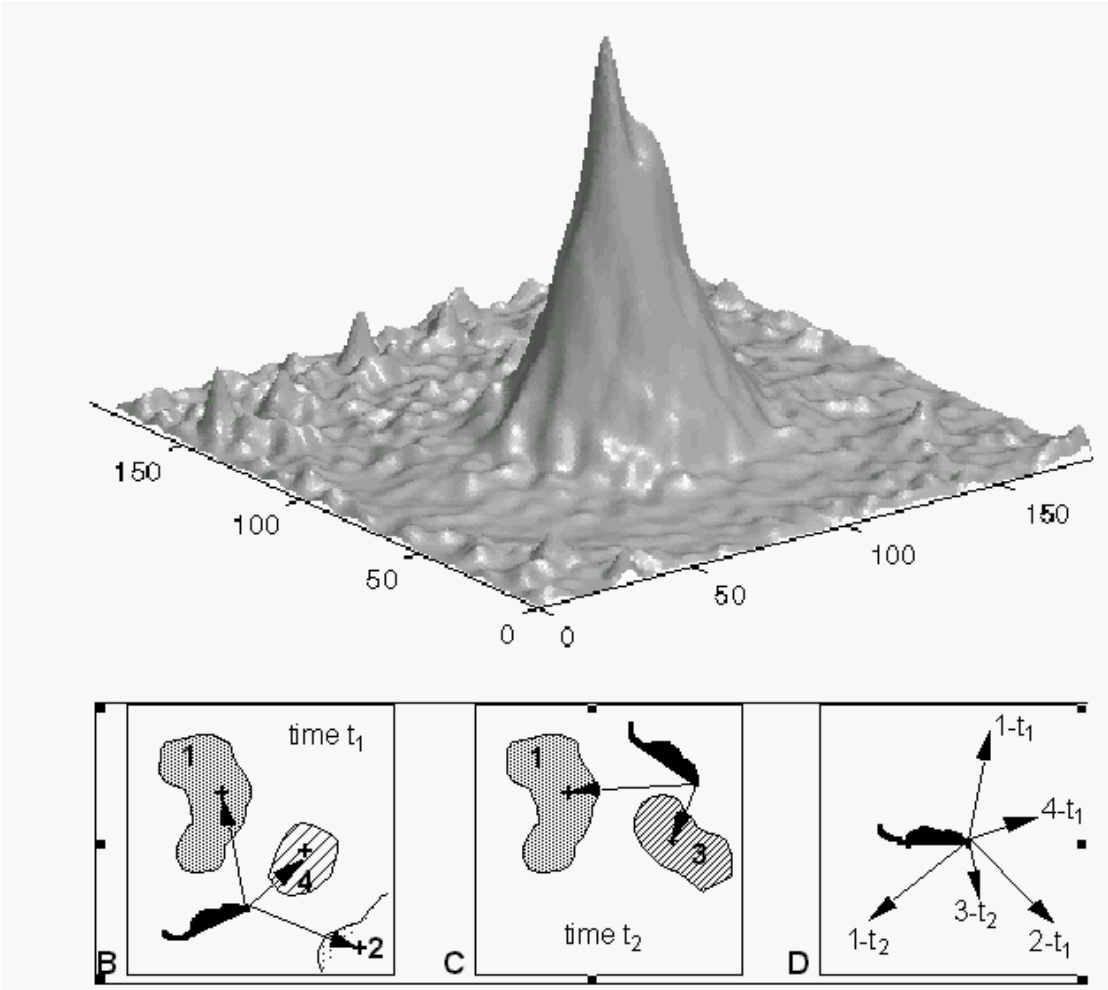
Neurons
Synapses
Neural networks
Patterns...

Пример: пространственная когнитивная карта



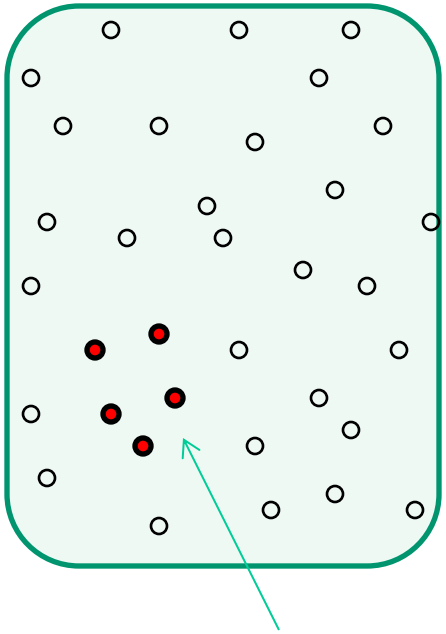


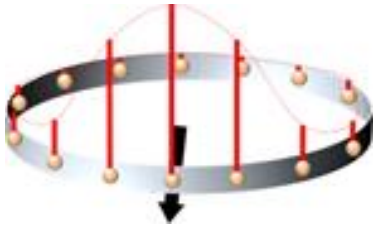
Пример: пространственная когнитивная карта



representation of experimental data (from Samsonovich & McNaughton,1997)

The concept

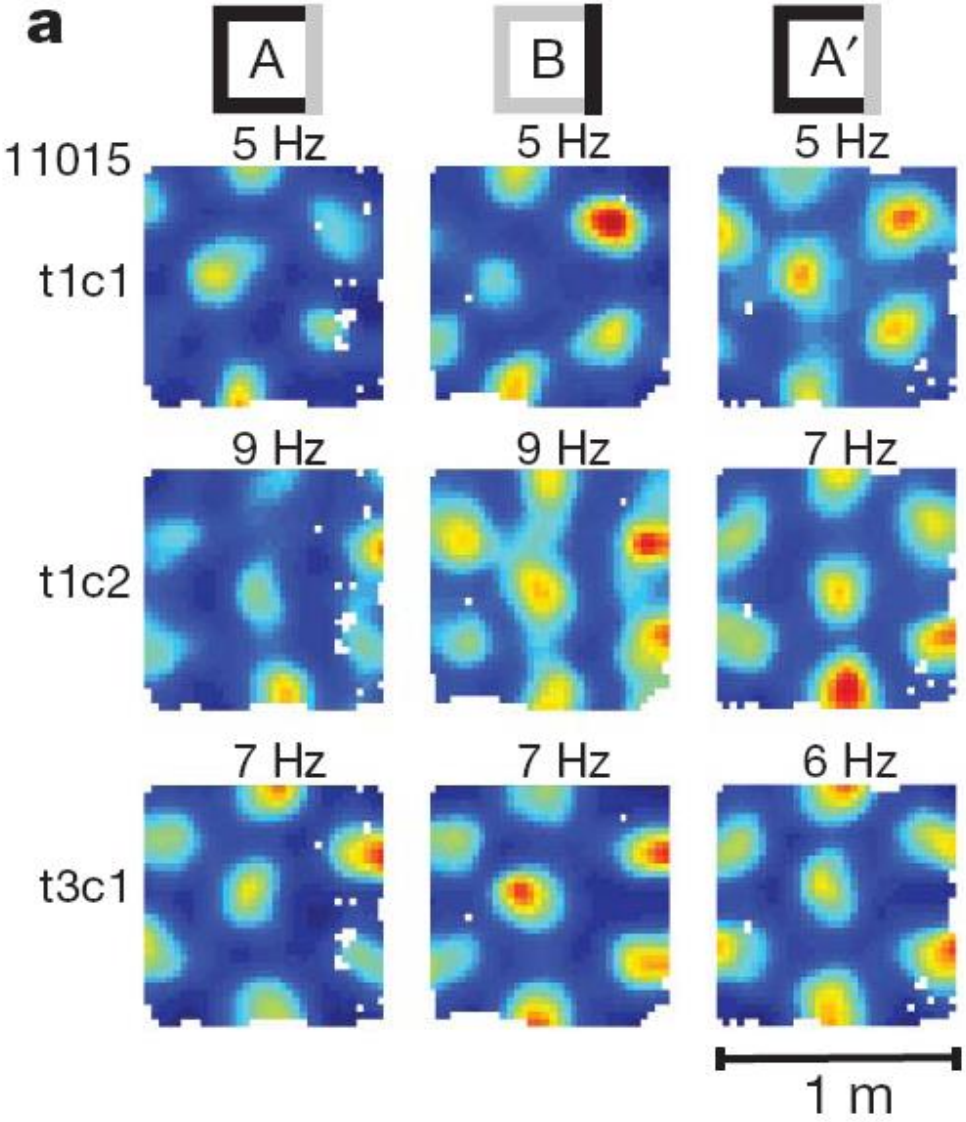




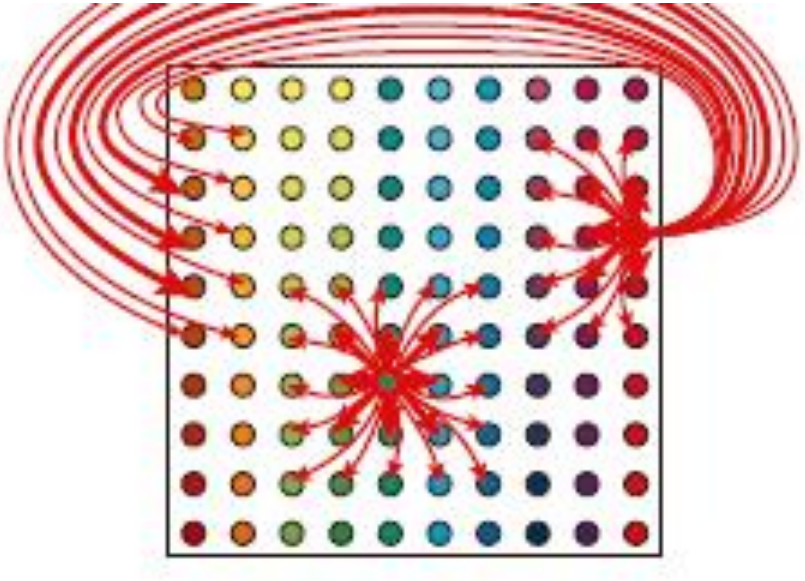
Head-direction cells

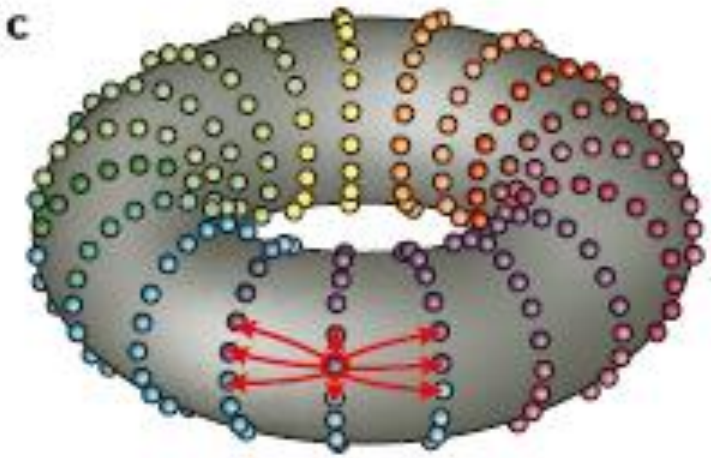
Пример: клетки решётки (grid cells)

- From Fyhn et al. Nature 2007



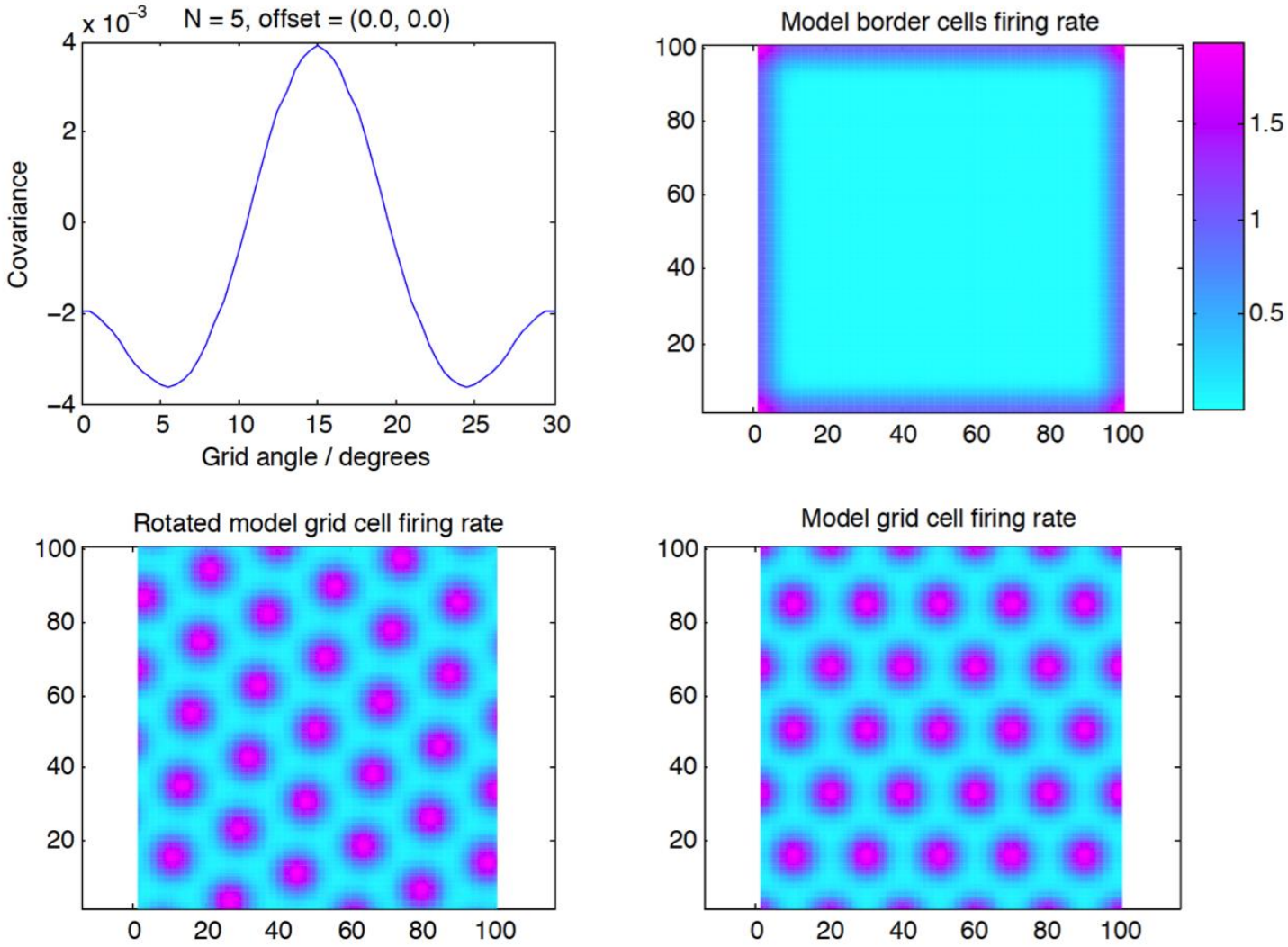
Пример: клетки решётки (grid cells)



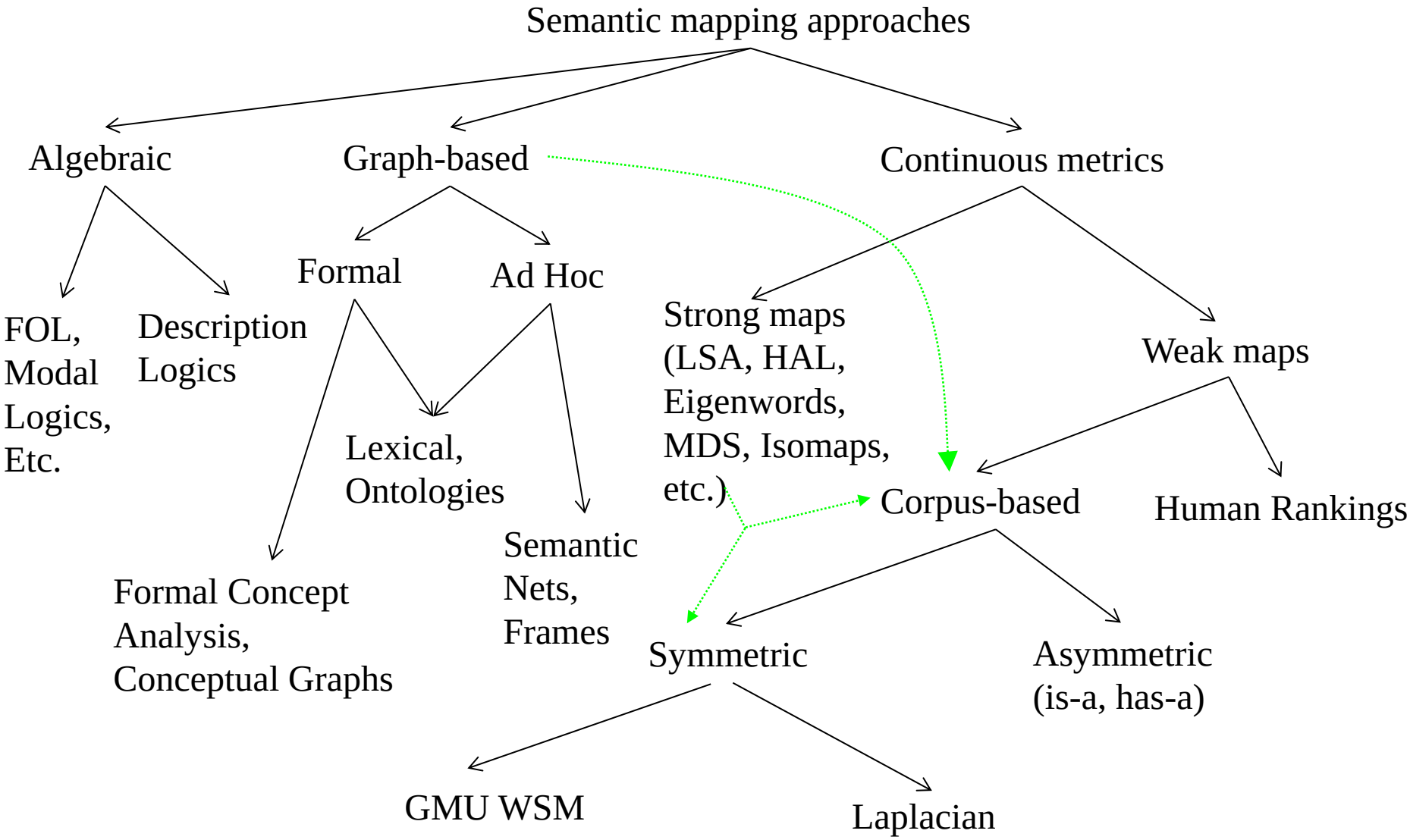


McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M. B. (2006). Path integration and the neural basis of the "cognitive map". *Nature Reviews Neuroscience* 7 (8): 663-678.

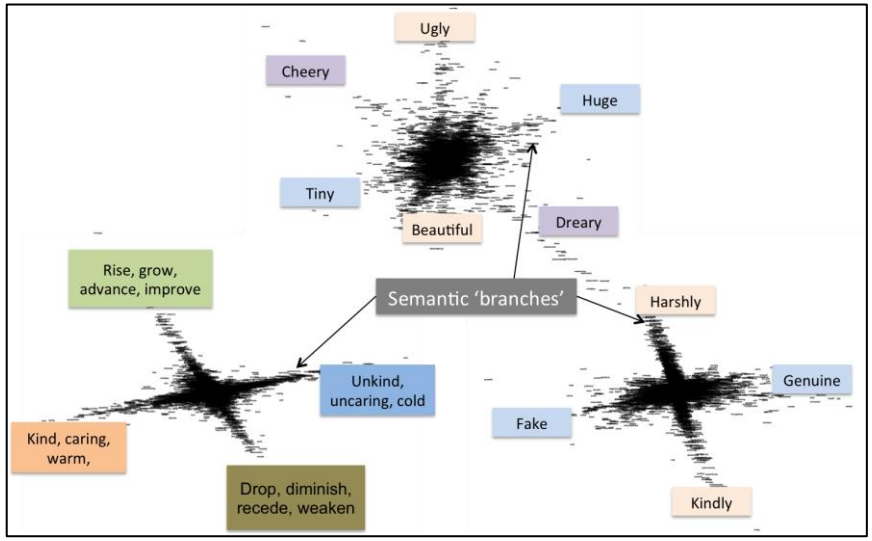
Пример: клетки решётки (grid cells)

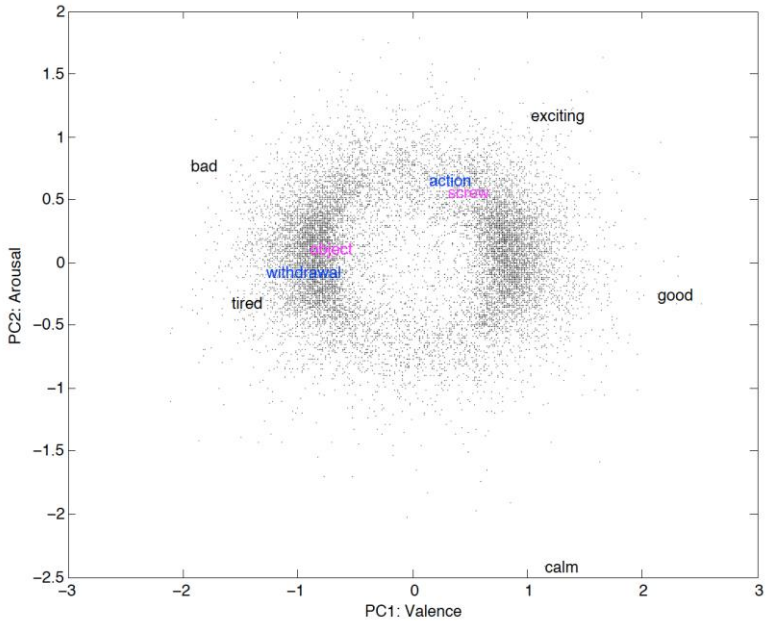


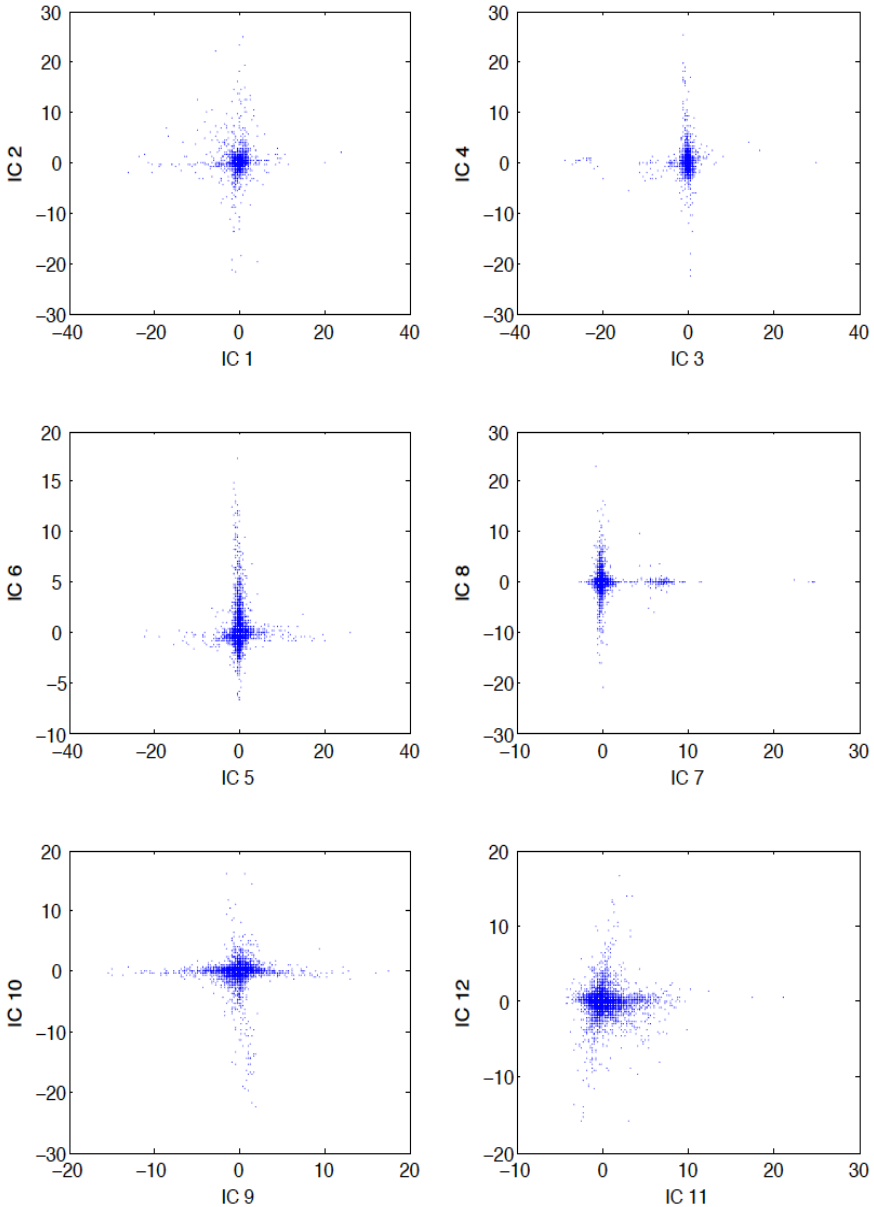
Пример: семантическое картирование



Laplacian embedding





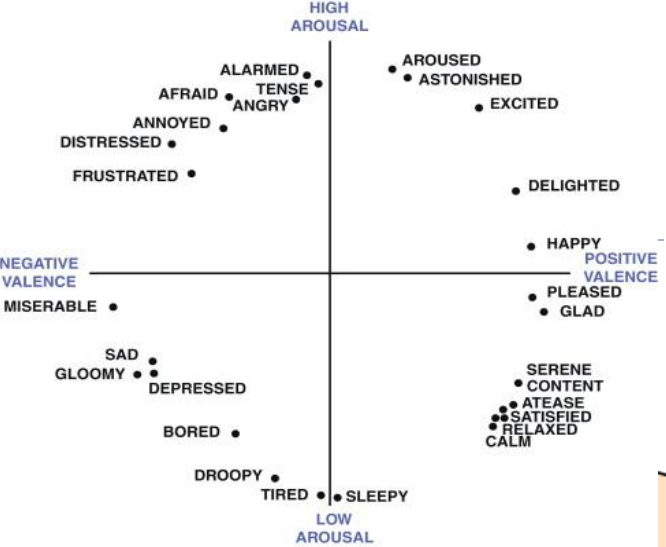


Laplacian embedding

little	unusual	sane	important	false	accept
small	strange	sensible	primary	fake	see
tiny	extraordinary	wise	crucial	unreal	regard
insignificant	peculiar	rational	fundamental	counterfeit	allow
minute	odd	reasonable	essential	sham	recognize
teeny	weird	balanced	significant	imitation	consider
wee	uncommon	intelligent	basic	artificial	have
unimportant	bizarre	lucid	serious	pretend	respect

large	usual	insane	unimportant	real	refuse
huge	familiar	daft	insignificant	genuine	reject
enormous	ordinary	nutty	minor	true	deny
massive	common	crackers	secondary	honest	automobile
big	normal	loony	trivial	natural	auto
immense	regular	nuts	subordinate	truthful	car
gigantic	standard	batty	petty	authentic	machine
vast	everyday	dotty	frivolous	sincere	ignore
colossal	customary	cracked	superficial	right	disregard
tremendous	native	loco	trifling	honorable	forbid

Emotion spaces



(Russell 1980)

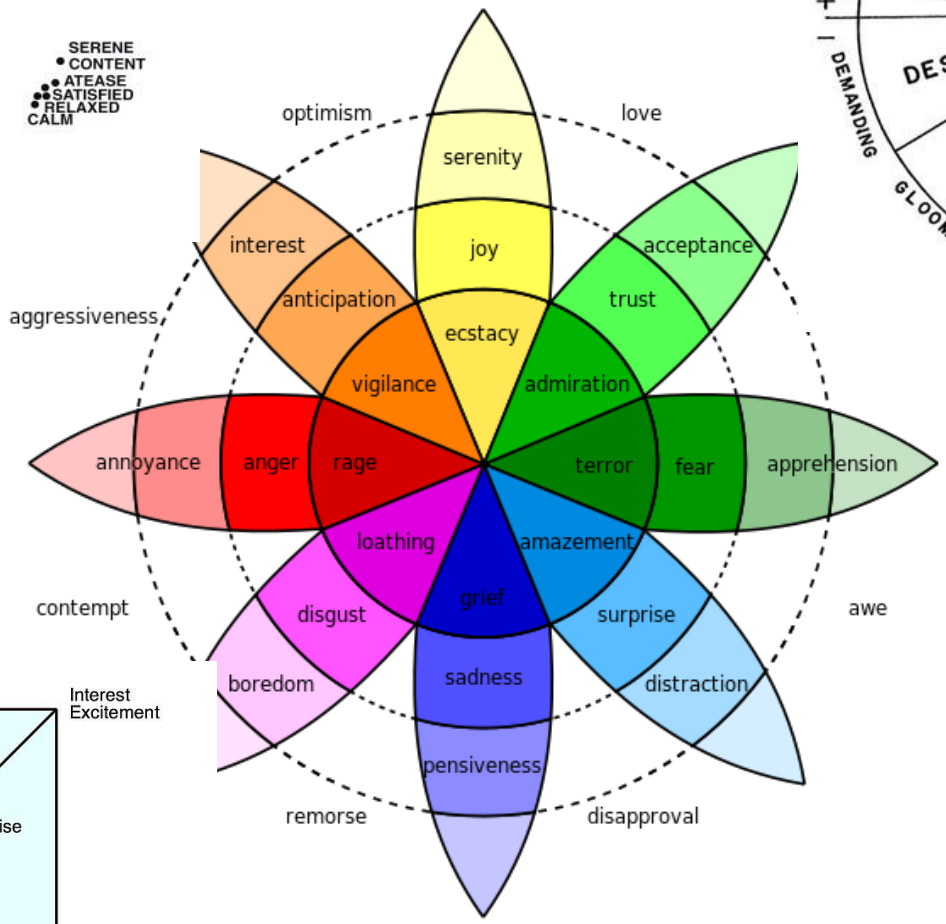
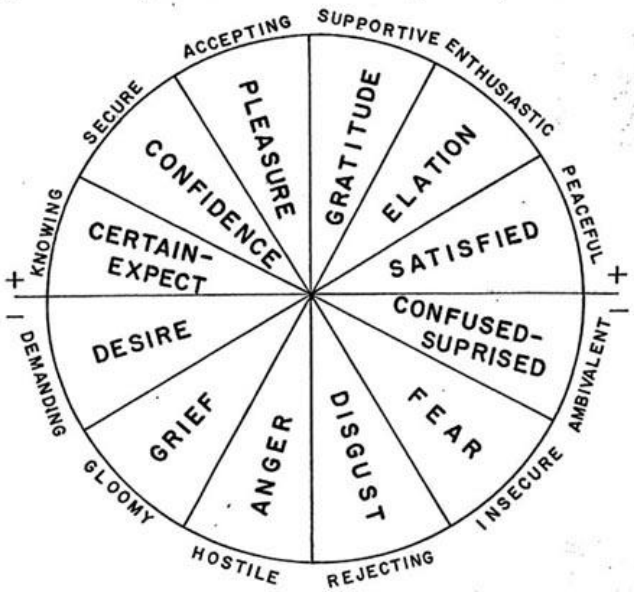
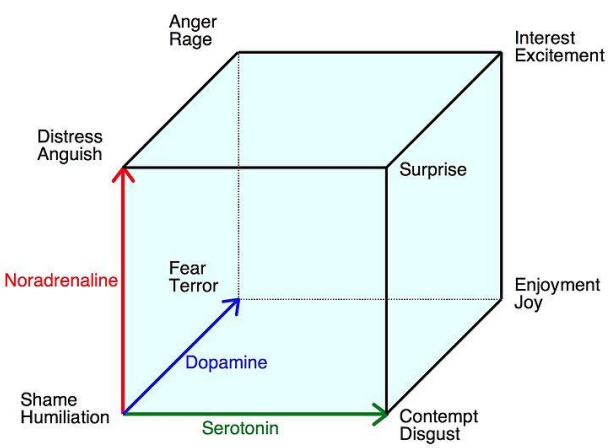


Figure 1: Personality Features derived from the Hexadyad Primary Emotions Model



← Plutchik

Cube of emotions (Lövheim 2012)

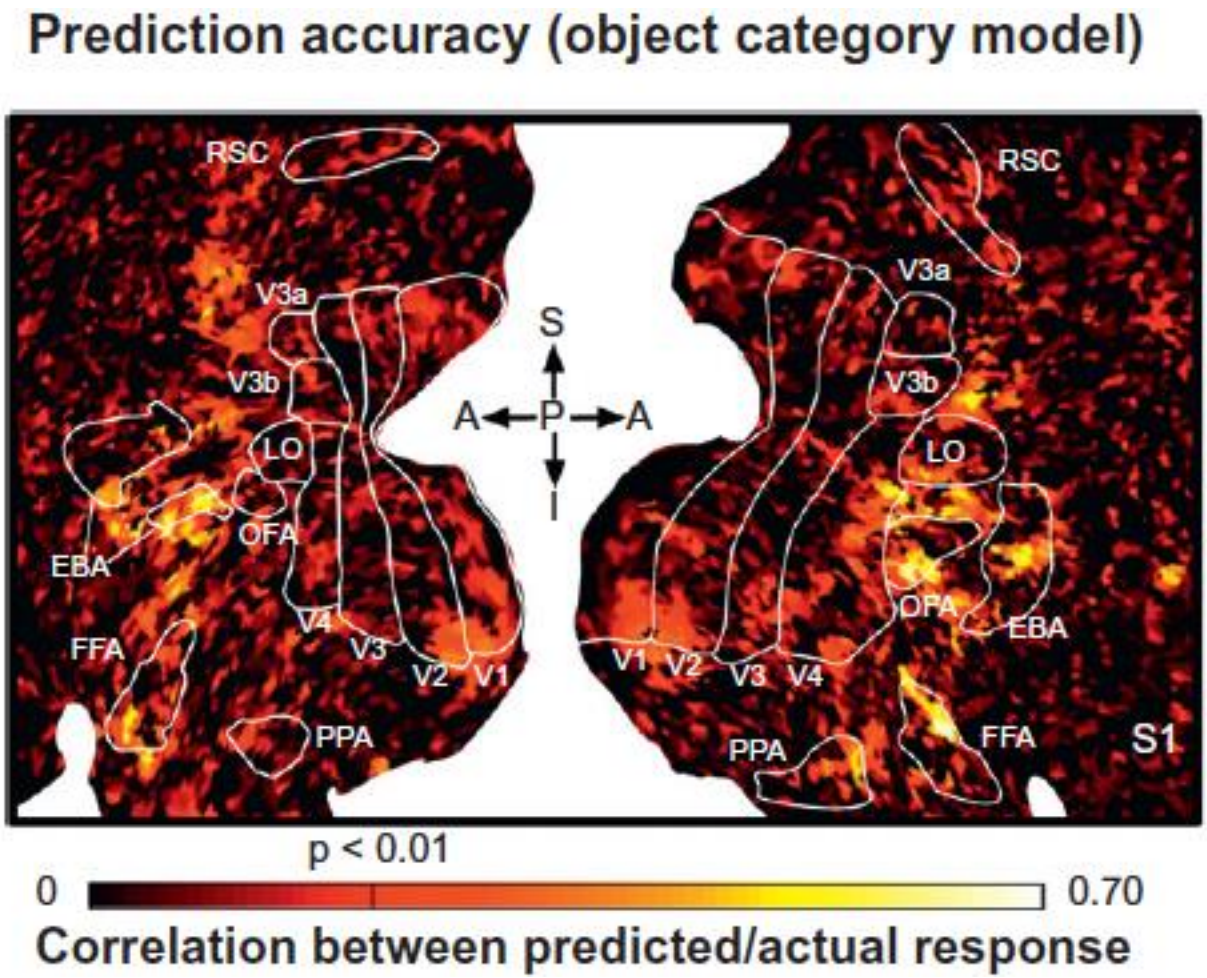


Other well-known models include:

- PAD (ANEW)
- EPA
- Semantic differential

Weak semantic map of emotional words

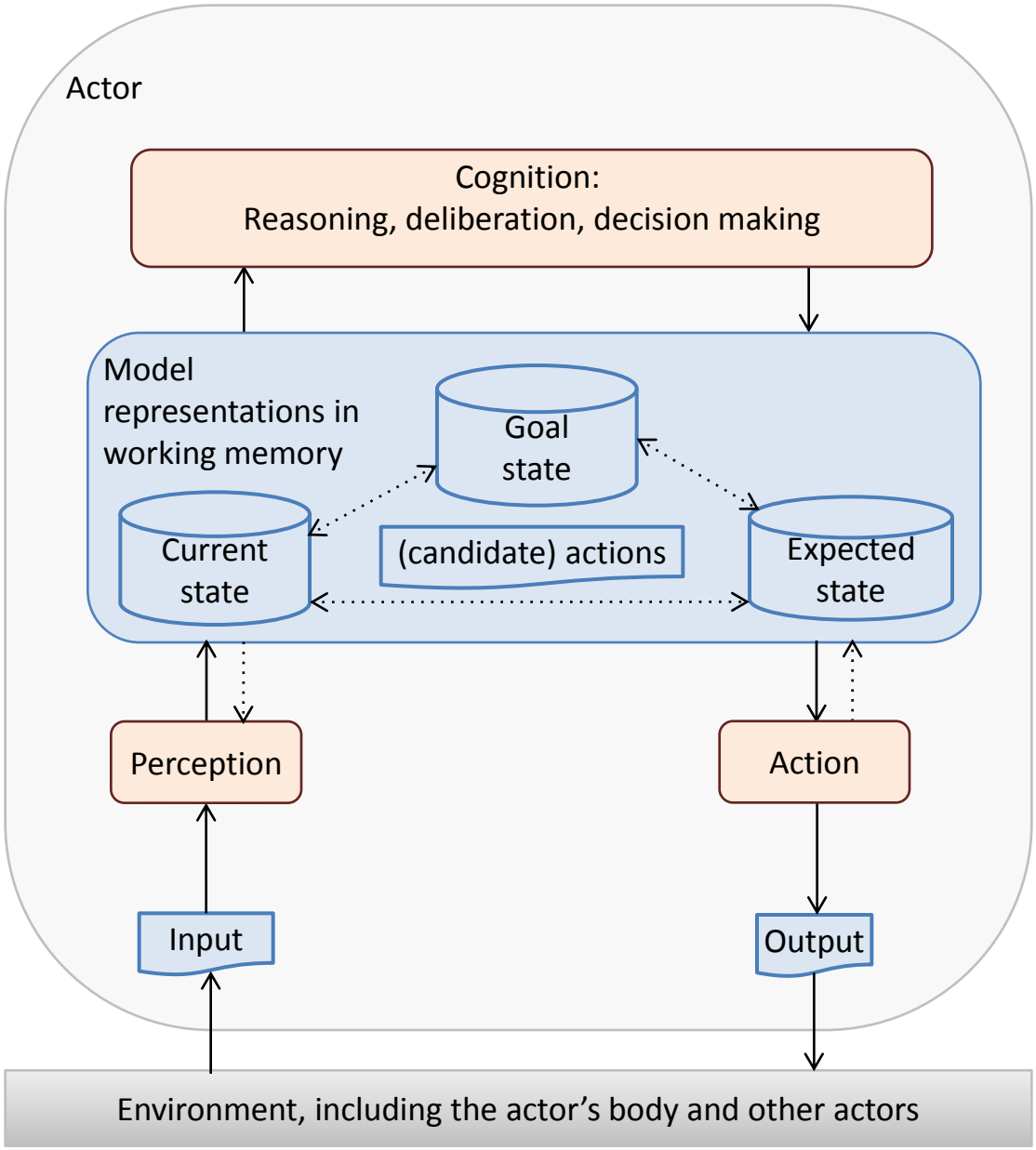
FROM: Thomas Naselaris, Dustin E. Stansbury b, Jack L. Gallant (2012). Cortical representation of animate and inanimate objects in complex natural scenes. *Journal of Physiology*.



BICA: definitions

- A *cognitive architecture* is a computational framework for designing intelligent agents.
- An agent, or *actor*, is a cognitive system embedded (not necessarily embodied) in a physical or virtual environment, such that it can perceive information and perform actions to satisfy its needs.
- A *cognitive system* is an information-processing dynamical system whose elements are functionally related to the semantics of the processed information.

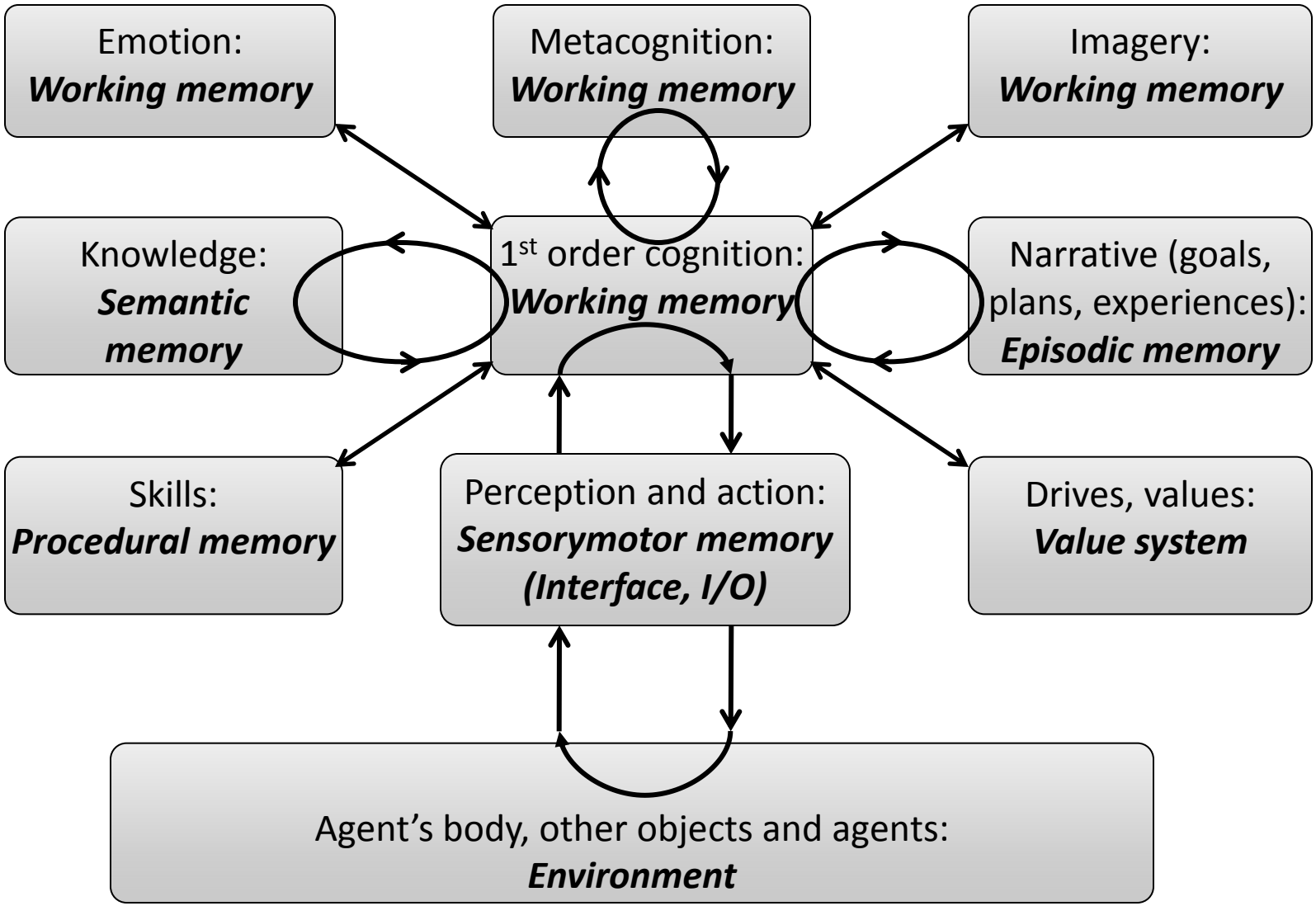
Basic cognitive cycle of a cognitive architecture



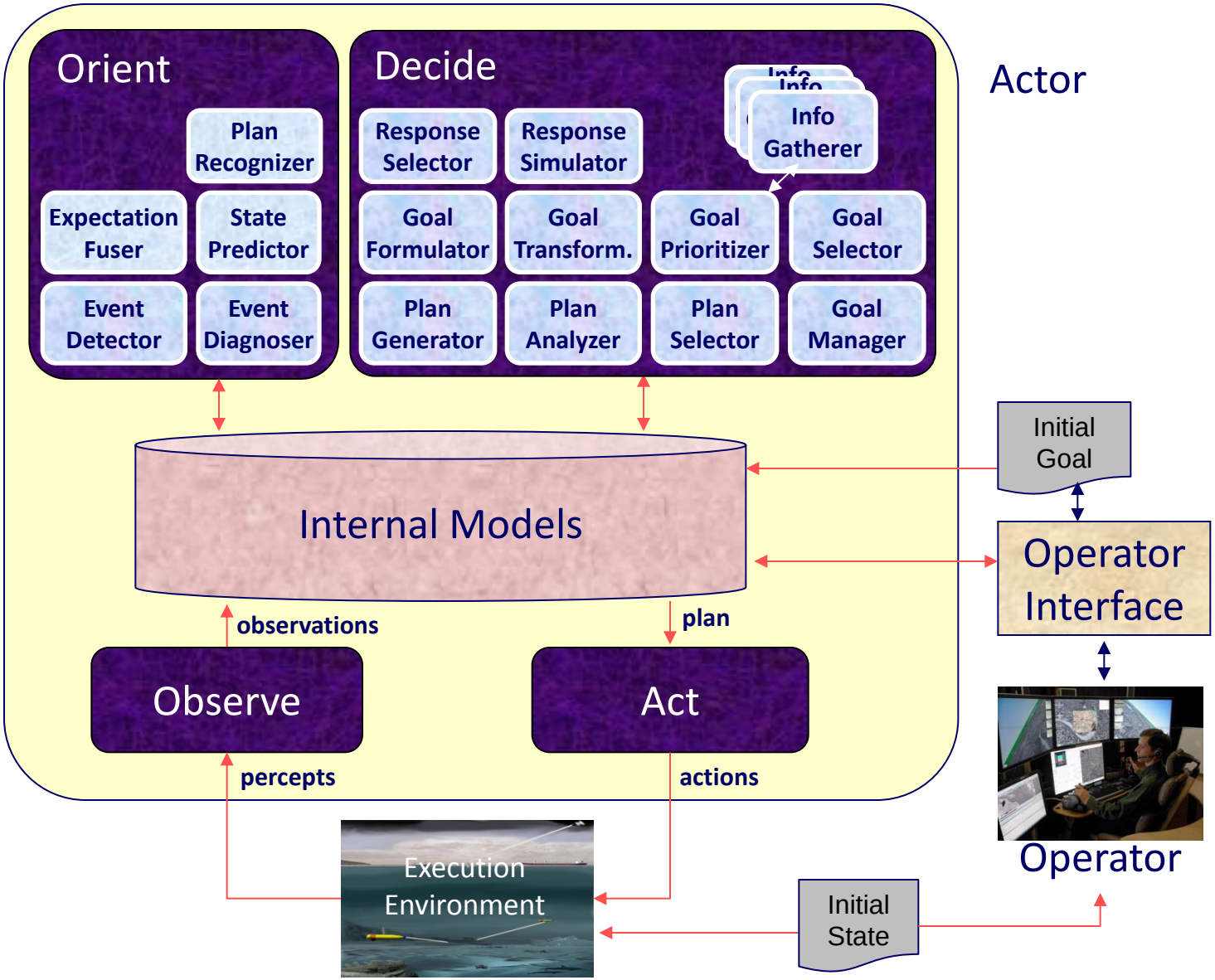
Hierarchy of cognitive architecture types

Architecture type and level	The agent is capable of
Metacognitive and self-aware (highest)	Modeling mental states of agents, including own mental states, based on a self concept
Reflective (high)	Modeling internally the environment and behavior of entities in it
Proactive, or deliberative (middle)	Reasoning, planning, exploration and decision making
Reactive, or adaptive (low)	Sub-cognitive forms of learning and adaptation
Reflexive (lowest)	Pre-programmed behavioral responses

Generic template for extended cognitive architectures



Generic goal reasoning model inspired by the OODA loop



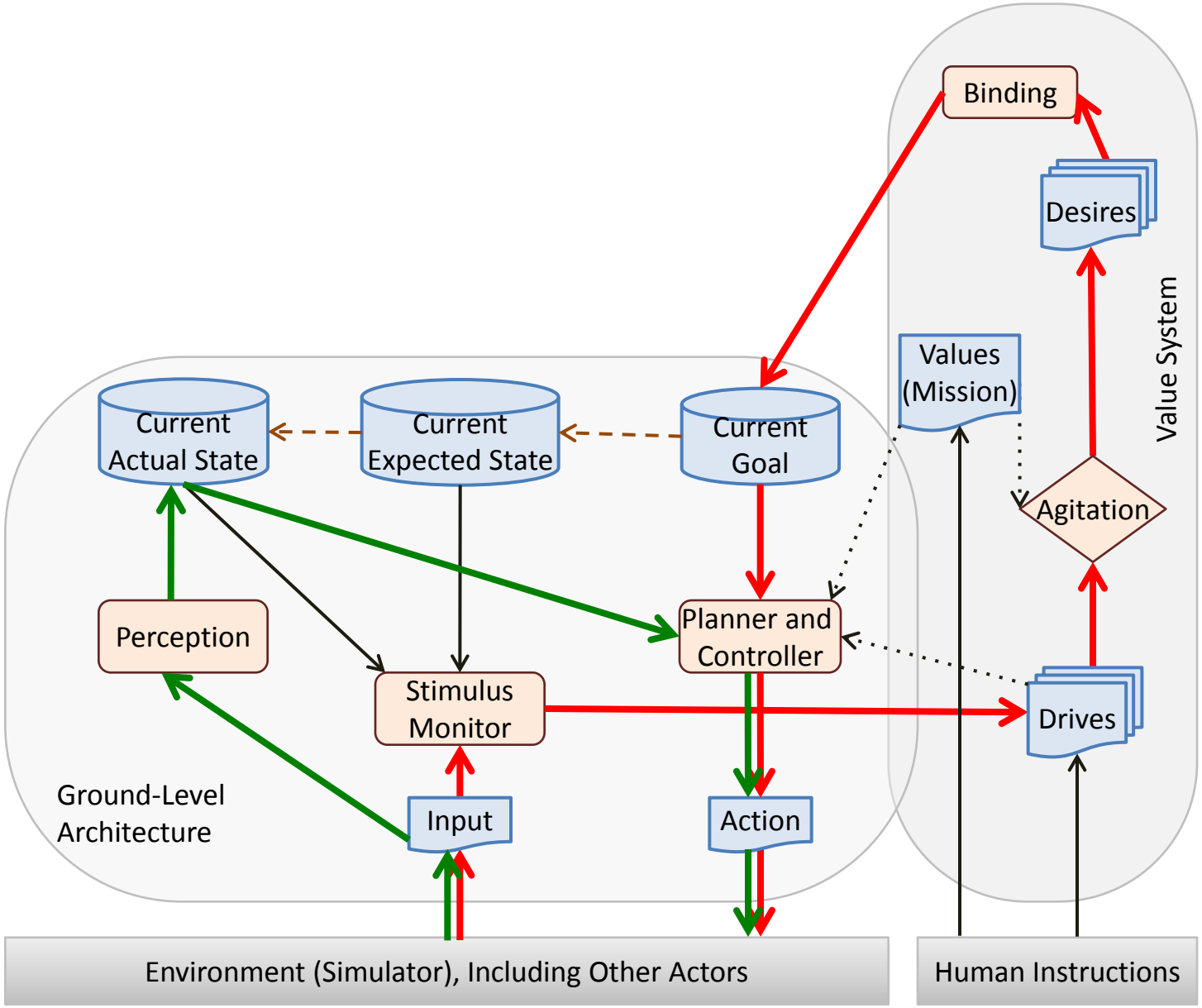
Definitions and formalization

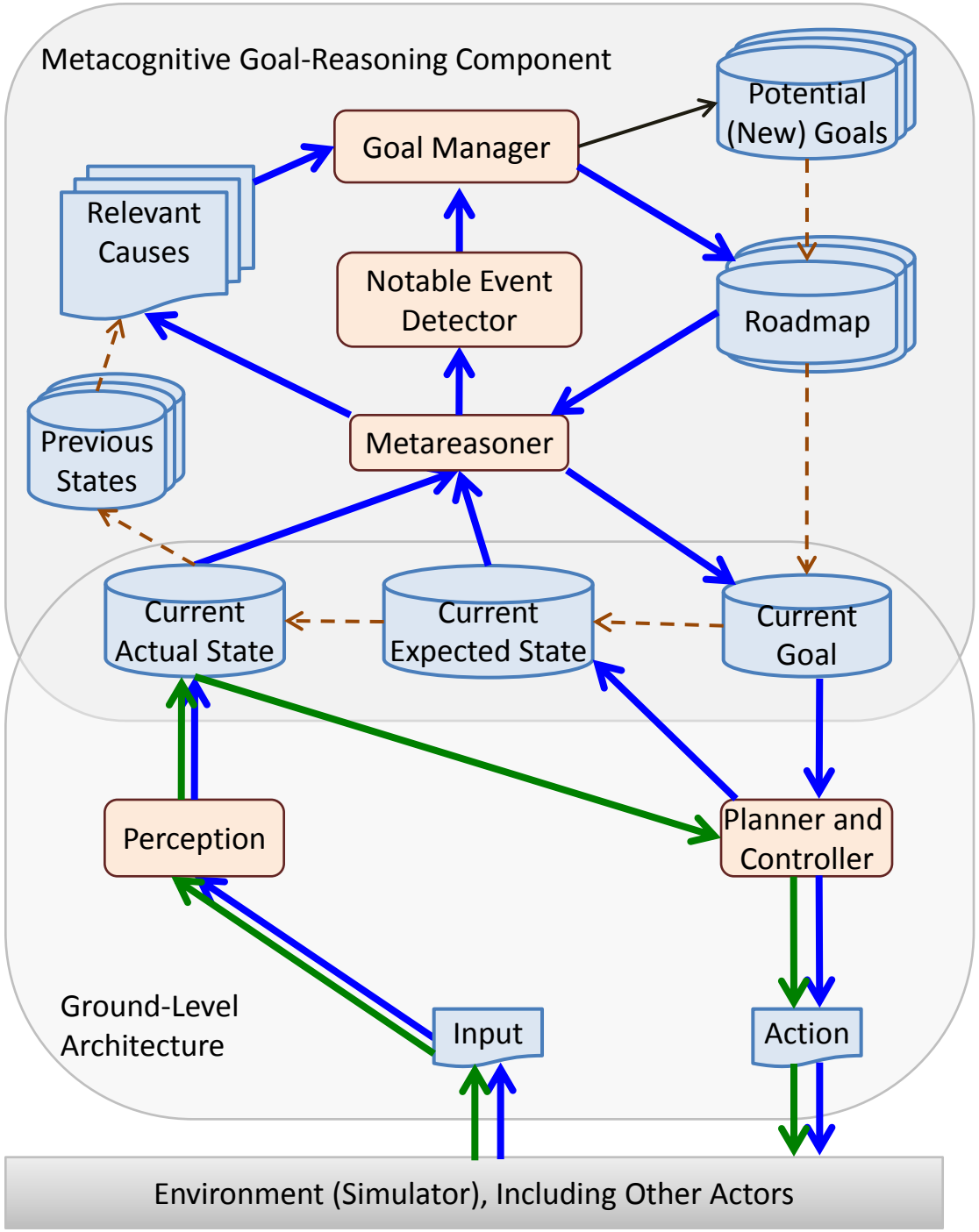
- The purpose of *goal reasoning* is to optimize the working narrative on the given cognitive map, following certain rules and policies. This can be done using a metacognitive cycle operating on the cognitive map, similarly to the base cognitive cycle operating on the working memory representation of the current state and the current goal.

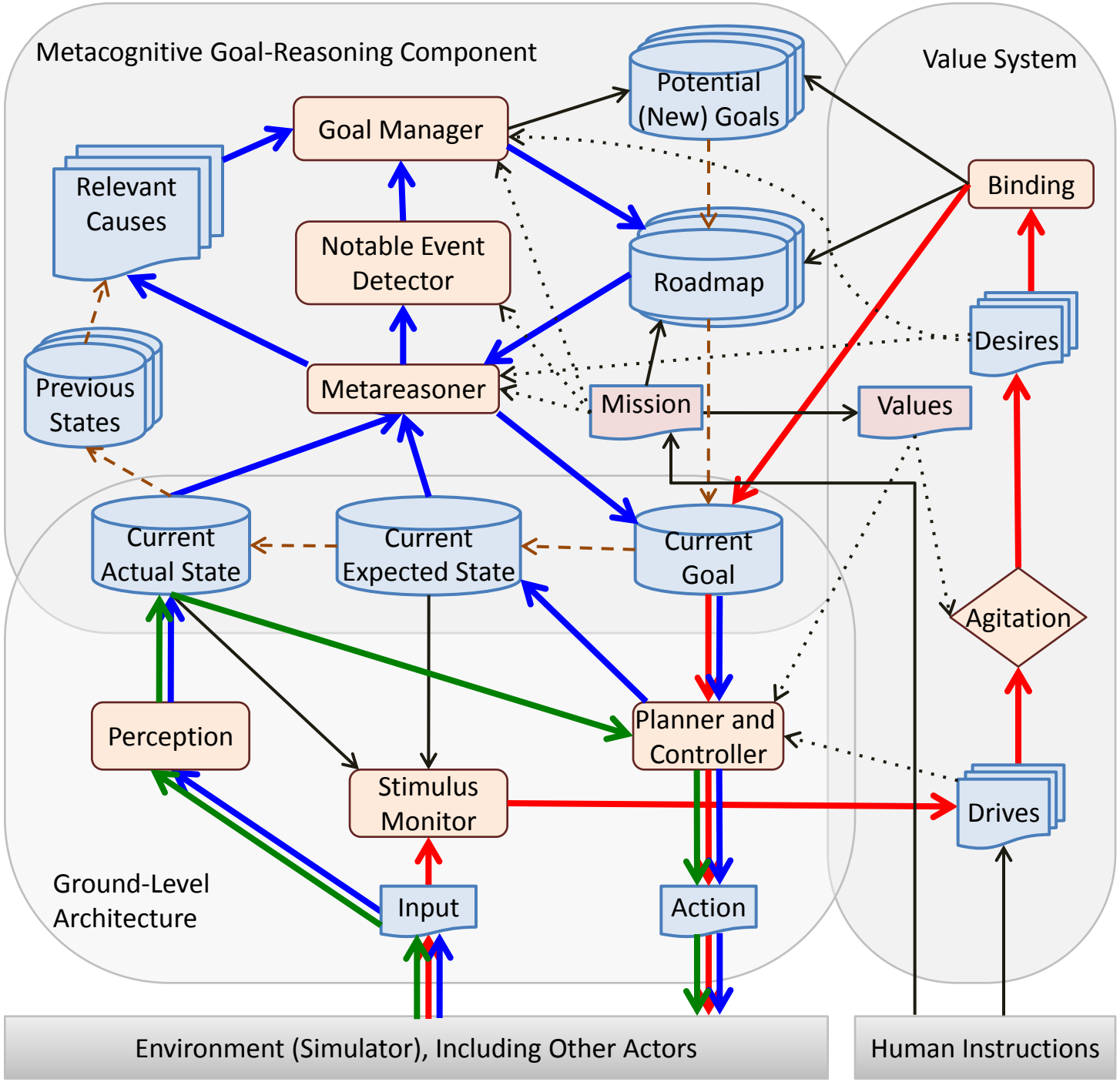
- **Goal reasoning cycle**
    - Operates on cognitive map
    - Triggered by notable phenomena
  - Detect notable phenomena that require attention
  - Update states based on notable phenomena
  - Update and re-evaluate narratives
  - Select the working narrative and identify the current goal in it
  - Return to basic cognitive cycle

Compare with: formulated → selected → expanded → committed → dispatched → evaluated → finished

Specific example: drive-driven desire-based goal activation




Metacognitive Goal-Reasoning Component
Potential (New) Goals
Goal Manager
Relevant Causes
Notable Event Detector
Roadmap
Previous States
Metareasoner
Current Actual State
Current Expected State
Current Goal
Perception
Planner and Controller
Ground-Level Architecture
Input
Action
Environment (Simulator), Including Other Actors


Metacognitive Goal-Reasoning Component
Value System
Potential (New) Goals
Goal Manager
Binding
Relevant Causes
Notable Event Detector
Roadmap
Desires
Previous States
Metareasoner
Mission
Values
Current Actual State
Current Expected State
Current Goal
Agitation
Perception
Planner and Controller
Stimulus Monitor
Drives
Ground-Level Architecture
Input
Action
Environment (Simulator), Including Other Actors
Human Instructions

Нарративный интеллект

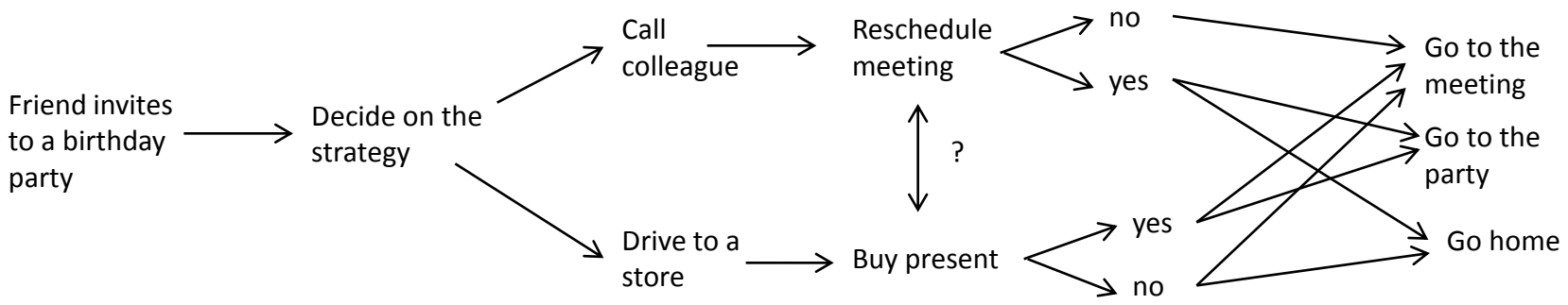
- Нарратив:
  - фабула, сюжет

- Нарративное планирование
- Нарративное целеполагание

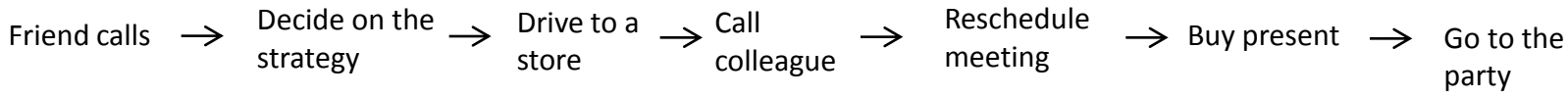
Уровень автора

Уровень актёра

A: “fabula”



B: “sjuzet”



IPOCL $(\langle S, B, O, L, C\rangle, F, \Lambda)$

The first parameter is a plan, with steps S, variable bindings B, ordering constraints O, causal links L, and frames of commitment C. F is a set of flaws (initially open conditions for each literal in the goal situation). Λ is a set of action schemata. Output is a complete plan according to Definition 6 or *fail*.

I. **Termination.** If O or B are inconsistent, fail. If F is empty and $\forall s \in S, \exists c \in C \mid s$ is part of c, return $\langle S, B, O, L, C\rangle$. Otherwise, if F is empty, fail.

II. **Plan Refinement.** Non-deterministically do one of the following.

- **Causal planning**
  1. **Goal selection.** Select an open condition flaw $f = \langle s_{\text{need}}, p\rangle$ from F. Let $F' = F - \{f\}$.
  2. **Operator selection.** Let s_{add} be a step that adds an effect e that can be unified with p (to create s_{add}, non-deterministically choose a step sold already in S or instantiate an action schema in Λ). If no such step exists, backtrack. Otherwise, let $S' = S \cup \{s_{\text{add}}\}$, $O' = O \cup \{s_{\text{add}} < s_{\text{need}}\}$, $B' = B \cup B_{\text{new}}$ where B_{new} are bindings (e.g., assignments of ground symbols to variables) needed to make s_{add} add e, including the bindings of s_{add} itself, and $L' = L \cup \{\langle s_{\text{add}}, e, p, s_{\text{need}}\rangle\}$. If $s_{\text{add}} \neq s_{\text{old}}$, add new open condition flaws to F' for every precondition of s_{add}.
  3. **Frame discovery.** Let $C' = C$.
     a. If $s_{\text{add}} \neq s_{\text{old}}$, non-deterministically choose an effect e of s_{add} or $e = nil$. If $e \neq nil$, construct a new frame of commitment c with internal character goal e and the character of s_{add}, let s_{add} be part of c, let $C' = C \cup \{c\}$, create a new open motivation flaw $f = \langle c\rangle$, and let $F' = F \cup \{f\}$.
     b. Let C'' be the set of existing frames of commitment that can be used to explain s_{add}. For all $d \in C''$, create an intent flaw $f = \langle s_{\text{add}}, d\rangle$ and let $F' = F \cup \{f\}$.
  4. **Threat resolution**
     - **Causal threat resolution.** Performed as in II.3 in the POCL algorithm (Figure 1)
     - **Intentional threat resolution.** For all $c_1 \in C'$ and $c_2 \in C'$, such that the character of c_1 is the same as the character of c_2, e_1 is the goal of c_1, and e_2 is the goal of c_2, if e_1 negates e_2, non-deterministically order c_1 before c_2 or vice versa and for all $s_1 \in c_1$ and all $s_2 \in c_2$, $O' = O' \cup \{s_1 < s_2\}$ or $O' = O' \cup \{s_2 < s_1\}$.
  5. **Recursive invocation.** Call IPOCL $(\langle S', B', O', L', C'\rangle, F', \Lambda)$.
- **Motivation planning**
  1. **Goal selection.** Select an open motivation flaw $f = \langle c\rangle$ from F. Let p be the condition of c. Let $F' = F - \{f\}$.
  2. **Operator selection.** Same as causal planning above, except $\forall s_i \in c, O' = O' \cup \{s_{\text{add}} < s_i\}$.
  3. **Frame discovery.** Same as for causal planning, above.
  4. **Threat resolution.** Same as for causal planning, above.
  5. **Recursive invocation.** Call IPOCL $(\langle S', B', O', L', C'\rangle, F', \Lambda)$.
- **Intent planning**
  1. **Goal selection.** Select an intent flaw $f = \langle s, c\rangle$ from F. Let $F' = F - \{f\}$.
  2. **Frame selection.** Let $O' = O$. Non-deterministically choose to do one of the following.
     - Make s part of c. Let s_{m} be the motivating step of c. $O' = O' \cup \{s_{\text{m}} < s\}$. For all $c_i \in C$ such that c_i is ordered with respect to c, then for all $s_i \in c_i, O' = O' \cup \{s_i < s\}$ or $O' = O' \cup \{s < s_i\}$. For each $s_{\text{pred}} \in S$ such that $\langle s_{\text{pred}}, p, q, s\rangle \in L$ and s_{pred} and s have the same character, create an intent flaw $f = \langle s_{\text{pred}}, c\rangle$ and let $F' = F' \cup \{f\}$.
     - Do not make s part of c.
  3. **Recursive invocation.** Call IPOCL $(\langle S, B, O', L, C\rangle, F', \Lambda)$.

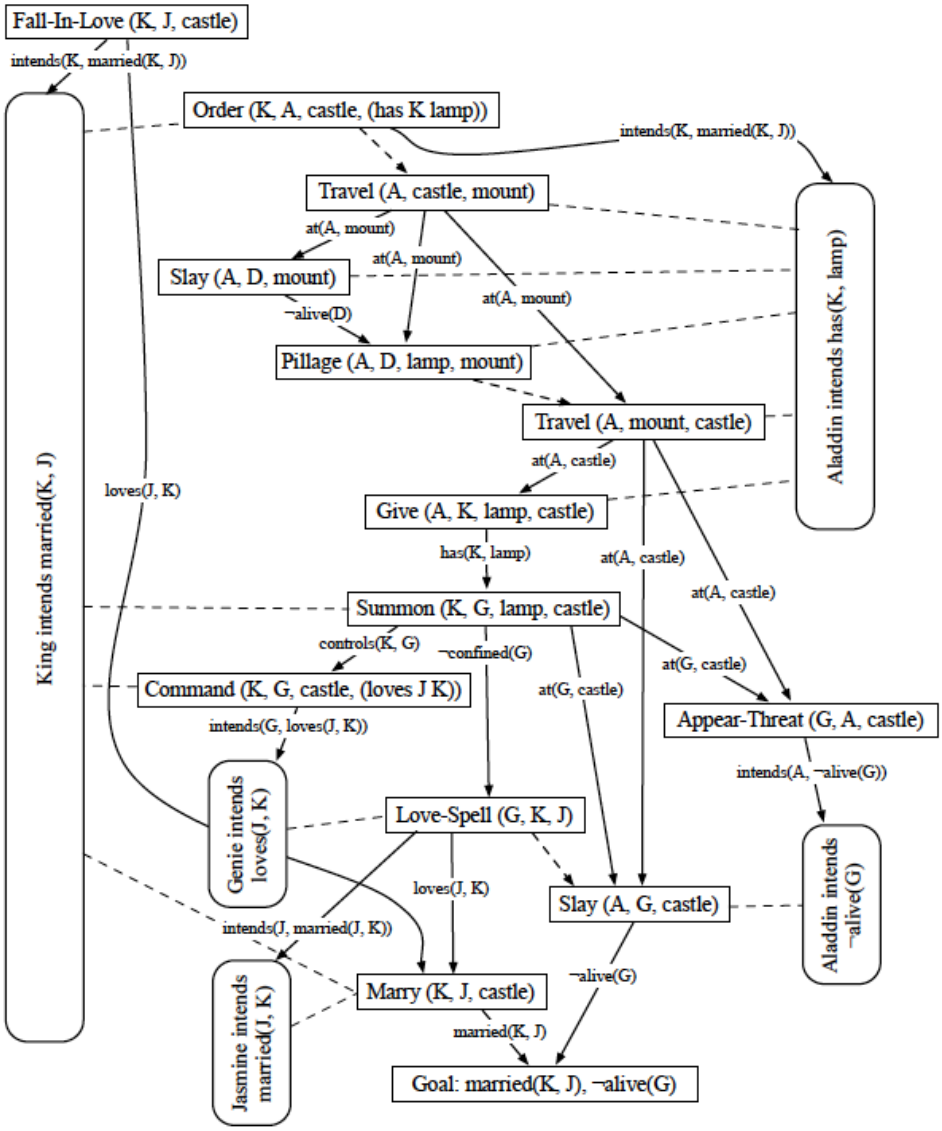


Figure 15: Fabula plan representation of the story used in the experimental IPOCL of the evaluation study.

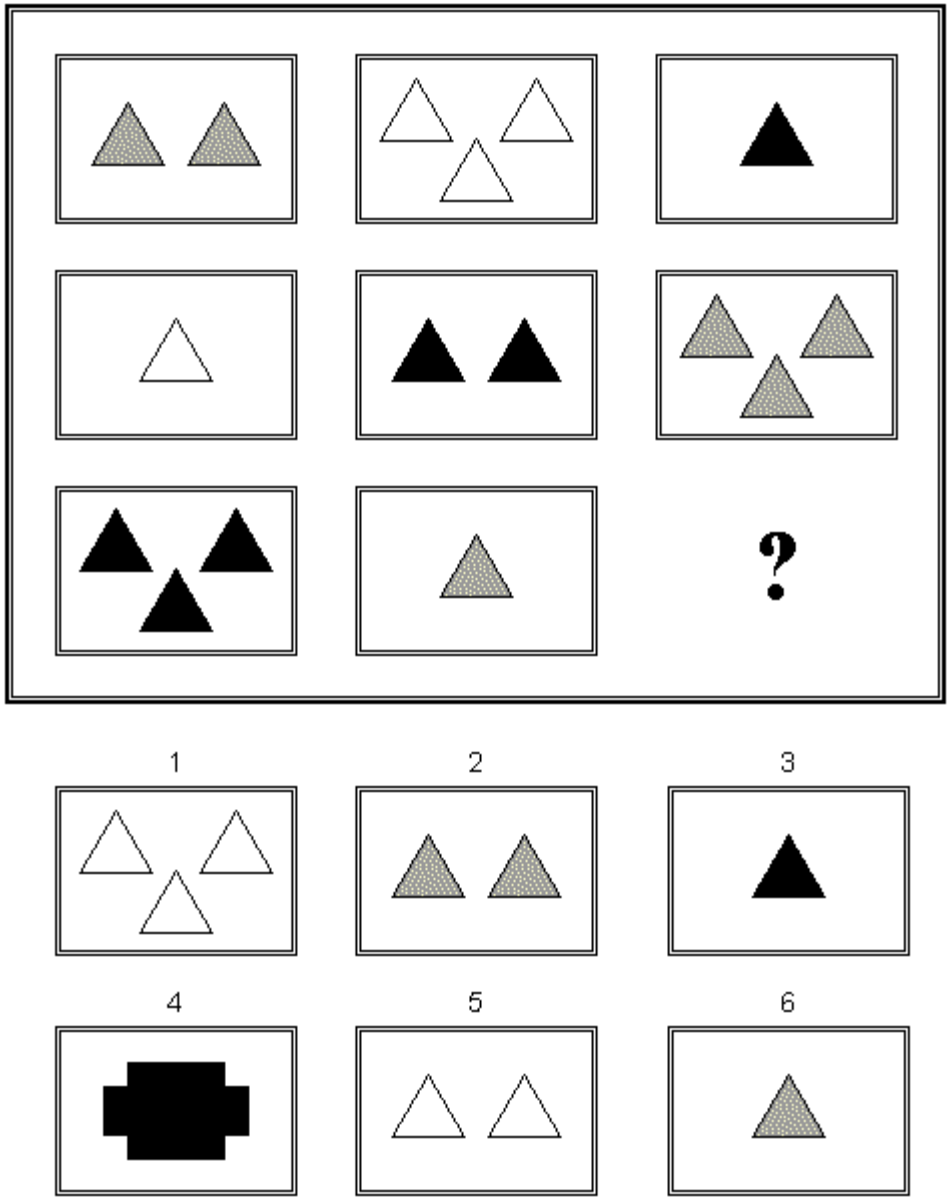
Возможные применения BICA

1. Emotionally intelligent social agents
2. Narrative-intelligent autonomous agents
3. Human-level learning agents
4. Fast & secure authentication system
5. BICA for elderly people and people with disabilities
6. BICA for artificial creativity

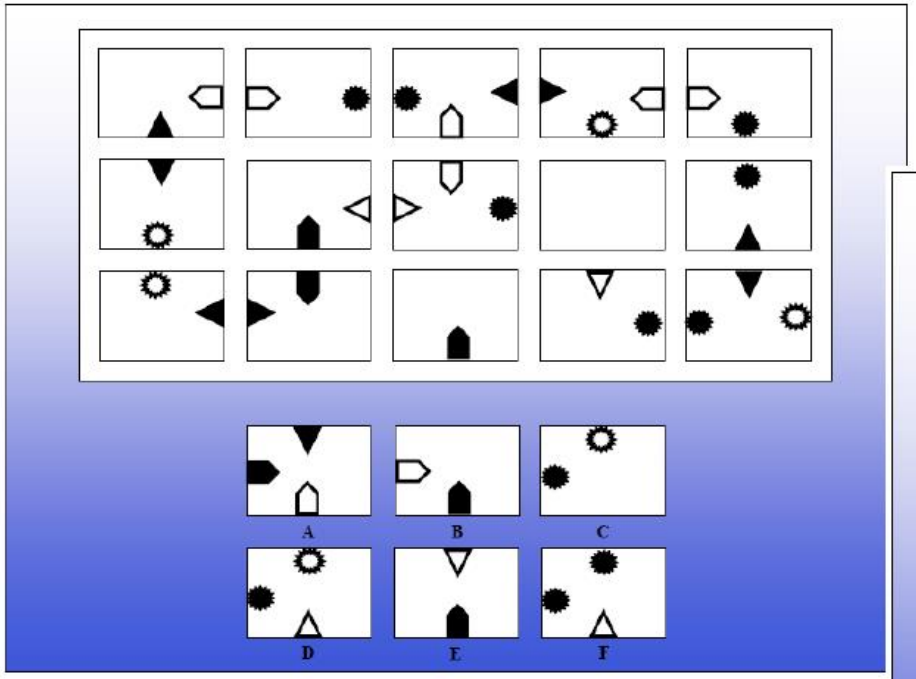
Искусственная креативность?

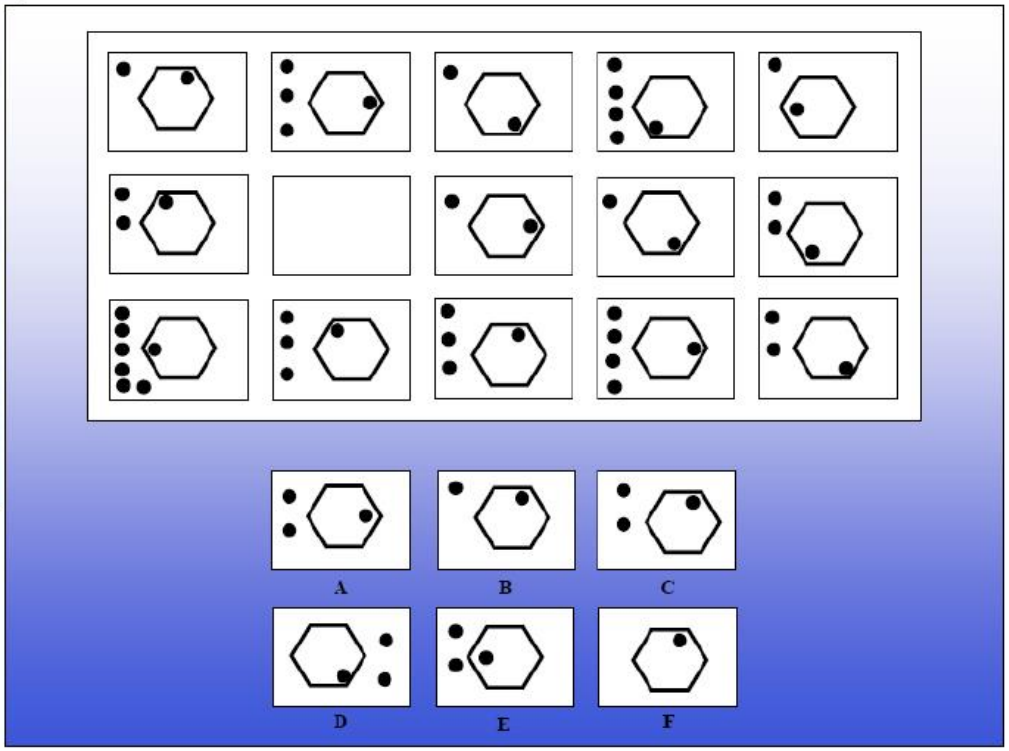
Fluid intelligence test

Which answer fits in the missing space to complete the pattern?

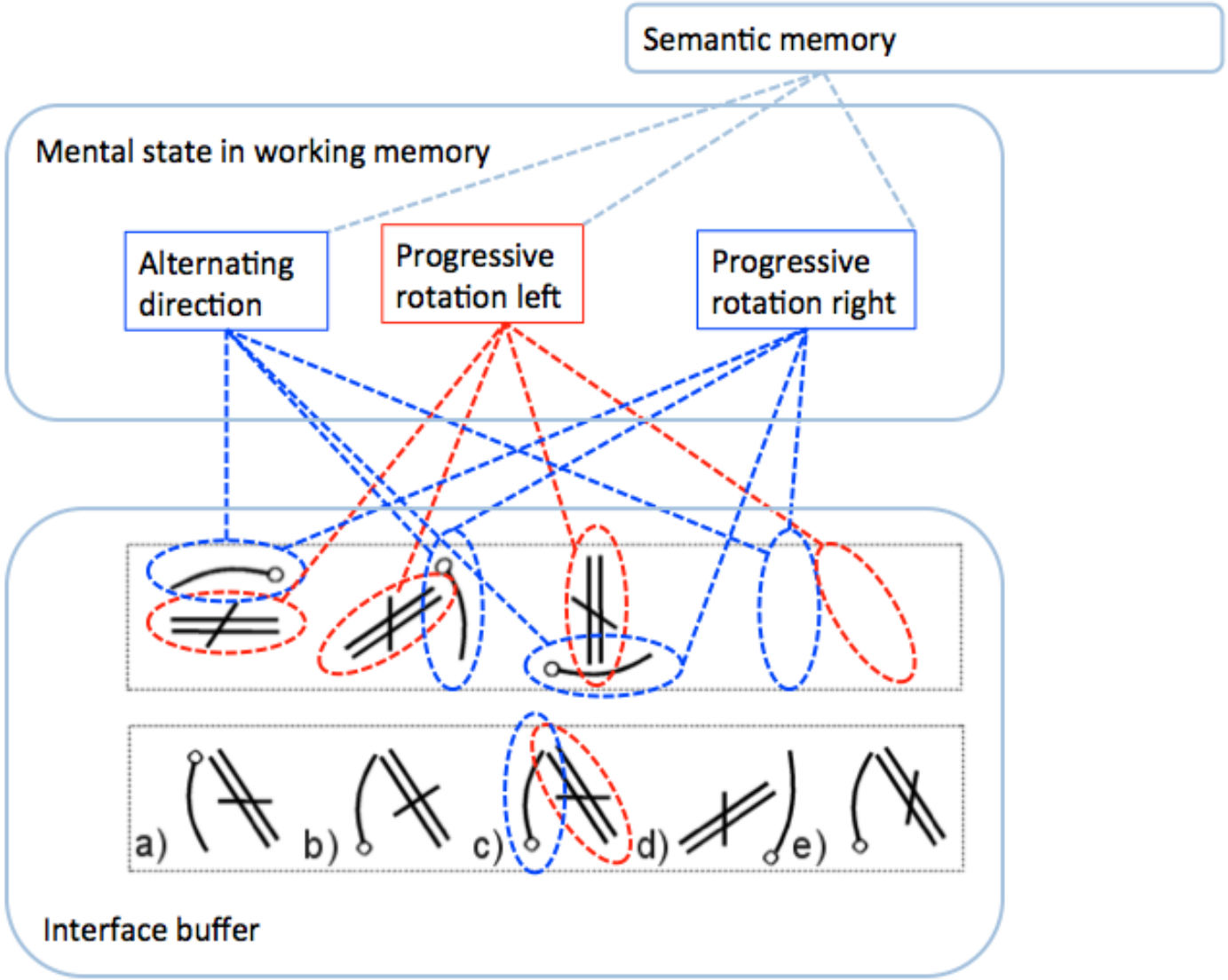


Bochumer Matrizentest (BOMAT®)





BICA model solving a test for fluid intelligence



- Четыре научных направления сводящиеся к одной задаче, и почему каждое из них само по себе не достаточно
- Разрыв между искусственным и естественным интеллектом
- Разрыв между железом и ПО
- Разрыв в понимании процессов лежащих в основе мышления
-
- Новая теоретическая парадигма объединяющая два уровня
- Почему не нужен уровень отдельных нейронов?
- Пример: пространственное мышление
- Пример: семантическое картирование
-
- Комплексный подход: Основа, Расширенные возможности
- Иерархия когнитивных архитектур

- Приложения
  - социальные агенты
    - эмоциональный, нарративный, обучаемый интеллект
  - надёжная аутентификация
  - помощь престарелым и людям с ограниченными возможностями
  - искусственное творчество

- Научное общество BICA Society:
- миссия координации усилий, конференция, журнал, видеопанели, таблица

Научное общество BICA Society

– Activities :

- The conference
- The journal BICA
- Videopanels
- Online Comparative Repository of Cognitive Architectures

Выводы:

- Для моделирования когнитивных функций человеческого мозга не требуется нейронный уровень
- Когнитивные карты – одна из основ интеллекта
- Биологически инспнрнрованные когнитивные архитектуры (BICA) выступают как объединяющая парадигма требующая комплексного подхода
- Ключевыми являются принципы социально-эмоционального и нарративного интеллекта (включая мета-мышление, автономный выбор целей, …), и человекоподобная обучаемость