Computational Concept and Architecture of Artificial Psyche

Anton Kolonin

akolonin@aigents.com
Telegram: akolonin

https://www.nsu.ru

https://aigents.com

https://agirussia.org
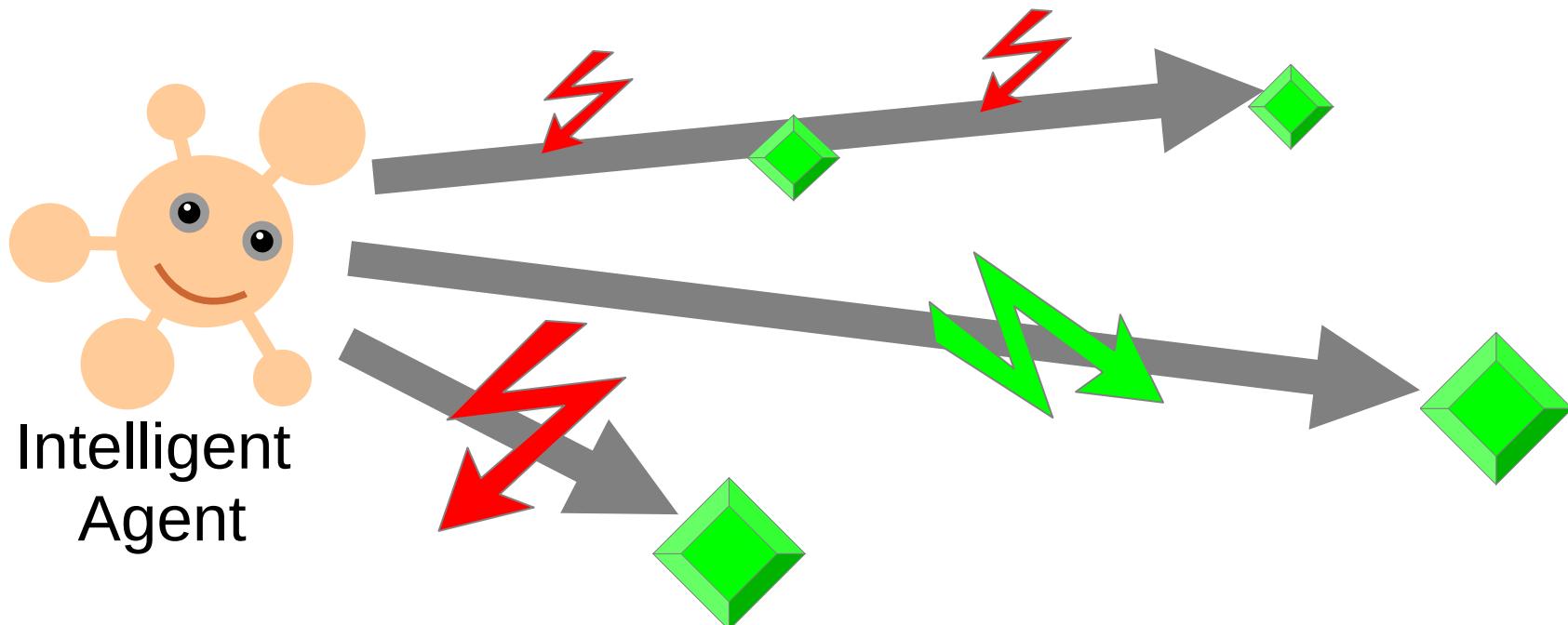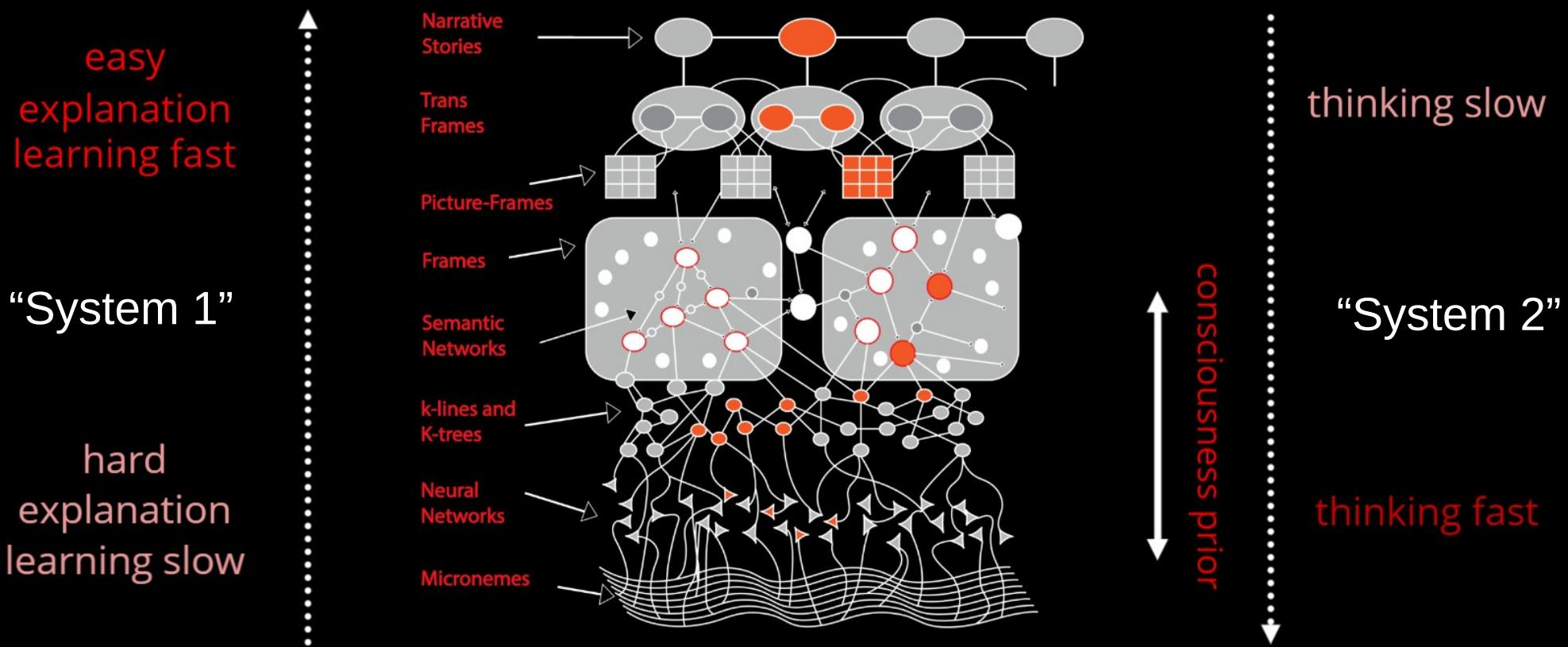
General Intelligence:

Reaching complex **goals** in different complex **environments**, using limited resources under uncertainty
(Ben Goertzel + Pei Wang + Shane Legg + Marcus Hutter)

Intelligent
Agent

2
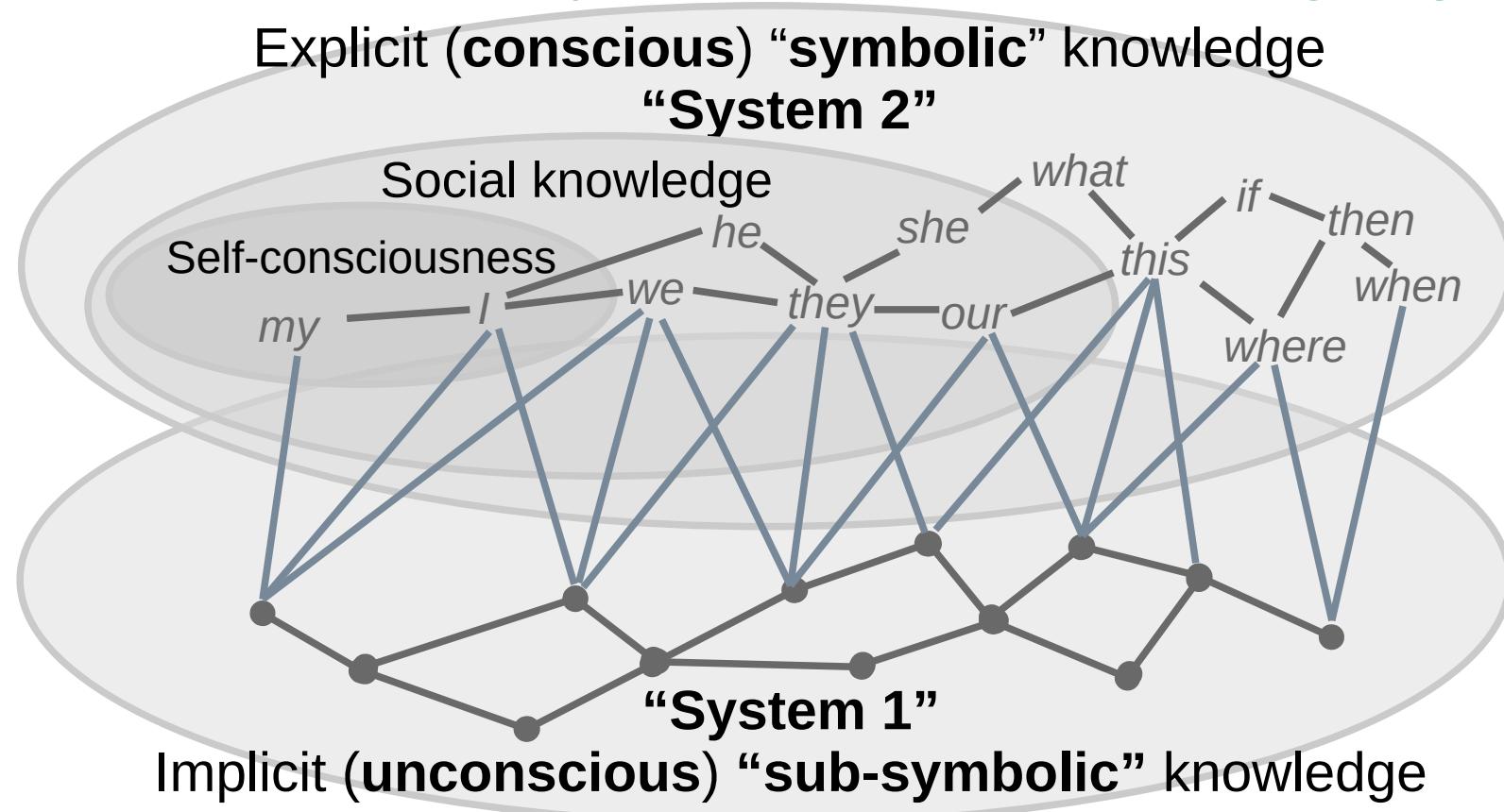
"Fast and Slow Thinking" – Daniel Kahneman
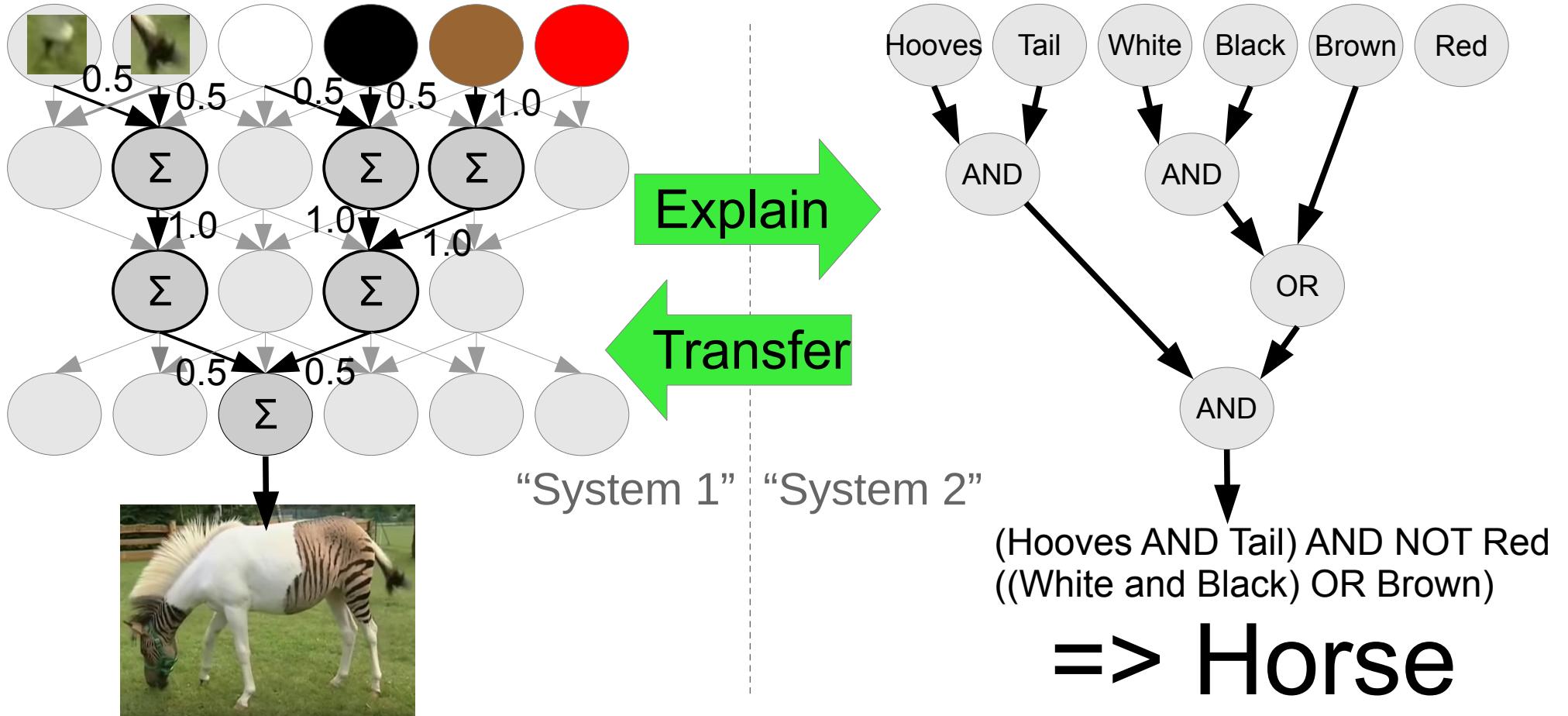
Medium: "neuro-symbolic" "knowledge graph"

Explicit (**conscious**) "**symbolic**" knowledge
"System 2"

Social knowledge

Self-consciousness

my *I* *we* *he* *they* *she* *our* *what* *this* *if* *then* *where* *when*

"System 1"

Implicit (**unconscious**) **"sub-symbolic"** knowledge

https://www.amazon.com/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374533555
https://amit02093.medium.com/atomspace-hyper-graph-information-retrieval-system-450cab9d751e

4
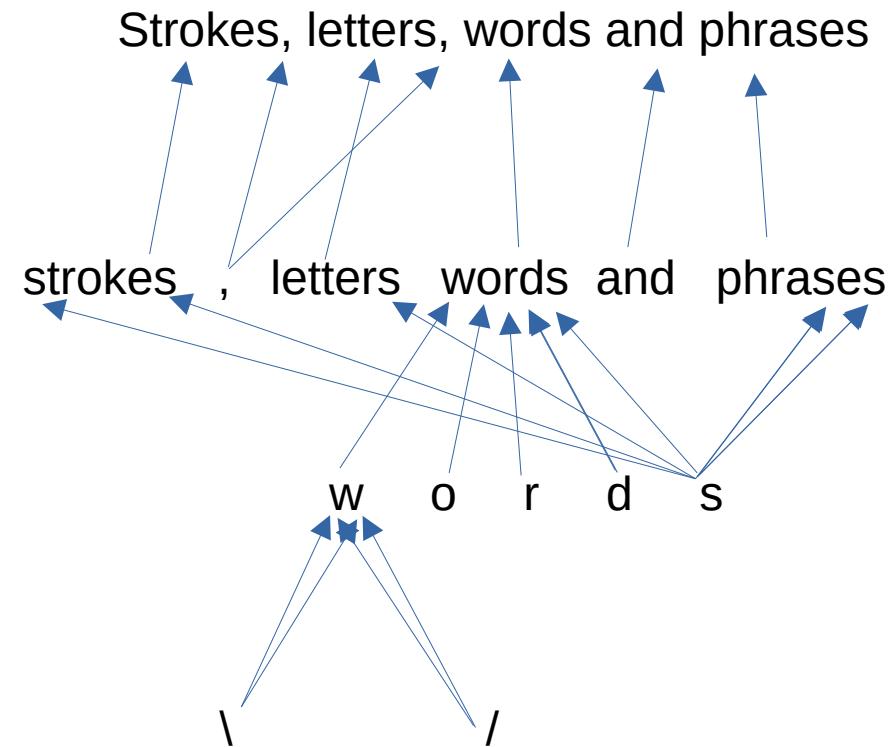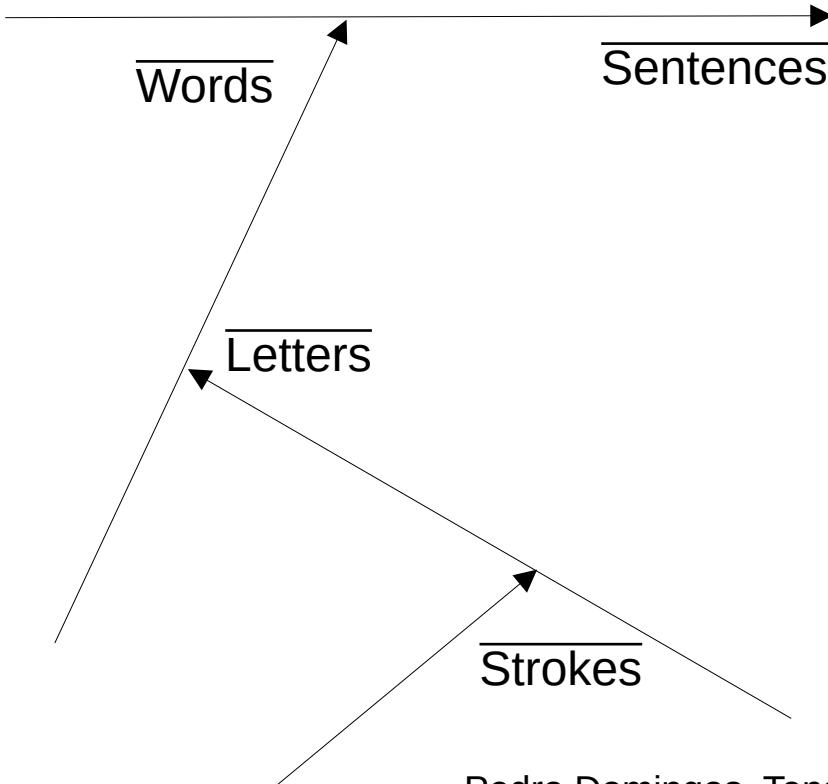
Neuro-Symbolic Integration for Interpretable AI

"System 1" "System 2"

(Hooves AND Tail) AND NOT Red
((White and Black) OR Brown)

=> Horse

5
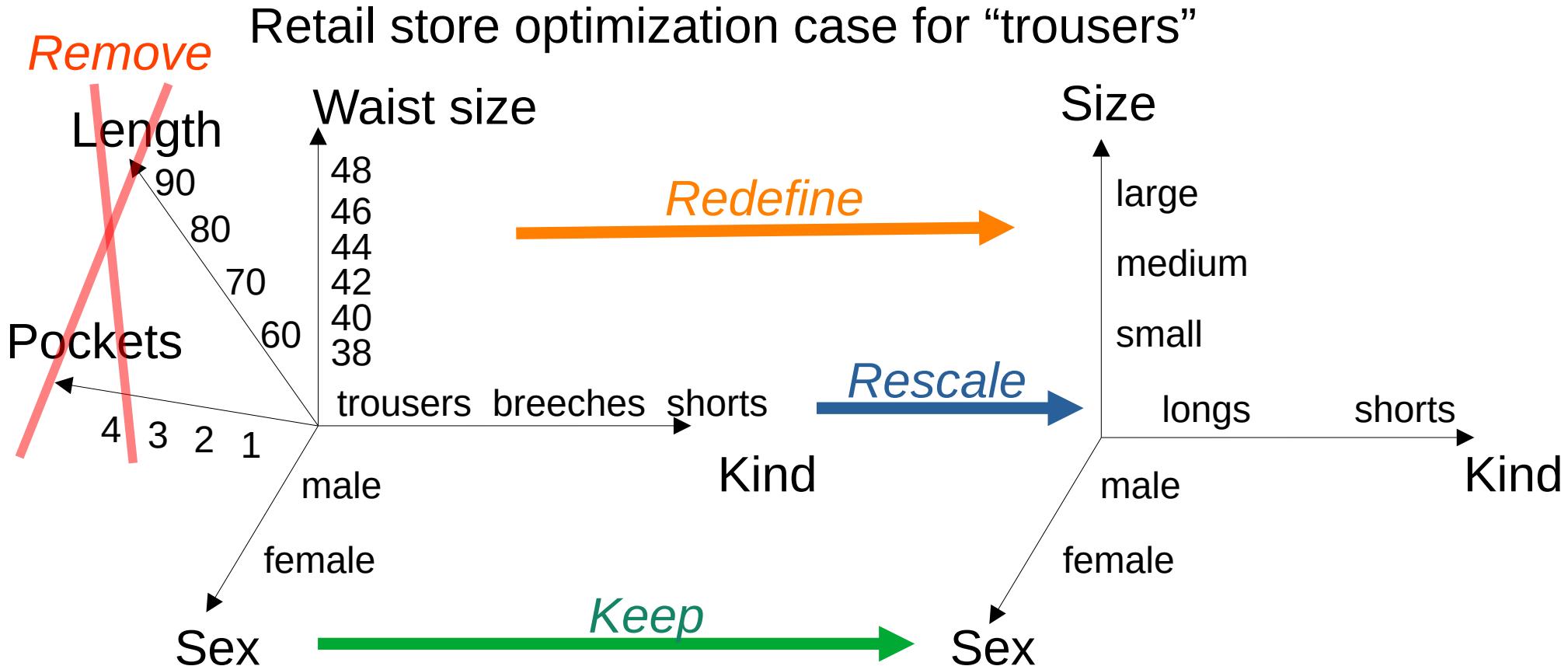
Functional equivalence of ~~neural network~~ tensor and graph (symbolic) models

Words

Sentences

Letters

Strokes

Strokes, letters, words and phrases

strokes , letters words and phrases

w   o   r   d   s

\   /
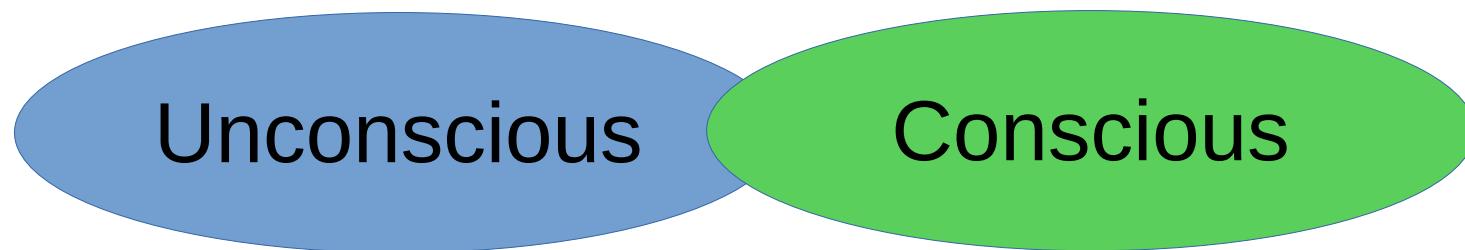
Pedro Domingos, Tensor Logic: The Language of AI
https://arxiv.org/pdf/2510.12269

Problem of dimensionality (reduction) and discreteness (increase)

Retail store optimization case for "trousers"

The psyche /ˈsaɪki/ is currently used to describe the totality of the human mind, conscious and unconscious.

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

Unconscious   Conscious

System = Intelligent agent

Psyche = Operating system

Intelligence = Decision making system

Subconsciousness (Intuition) "System 1" ("Fast")	Consciousness (Reasoning) "System 2" ("Slow")

Emotions → *Motivation* → *Expectations*

Space of states

Sensations	Needs	Actions

Decisions

Perceptions

Sensors

Actuators

Sensations

Actions

Outer world = Operational environment
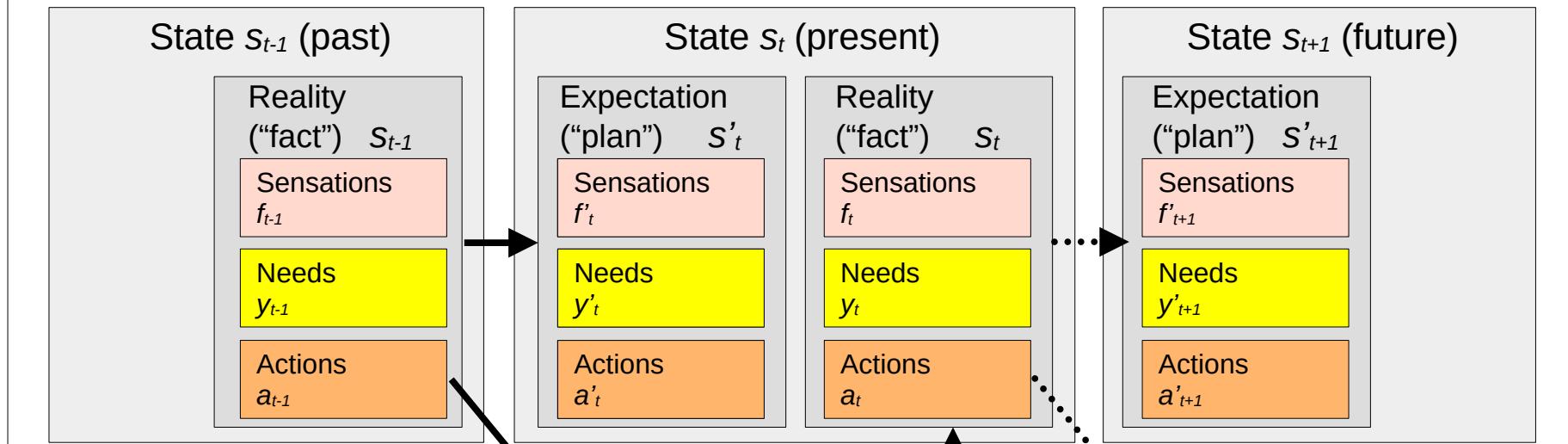
Anton Kolonin & Vladimir Kryukov, Computational Concept of the Psyche, Neuroinformatics-2025 https://arxiv.org/pdf/2509.07009

9

Psyche = Operating system

Intelligence = Decision making system

Models s ("invariants") of states with utilities U and probabilities P of transitions

$$U(\{s_\tau\}_{\tau \in \{-T,-1\}}, s'_0) = L(\ x \cdot (y_t - y_{t-1}),\ 1-(s'_t - s_t),\ 1-E(a_{t-1}))\qquad s'_t = argmax_s(U(\{s_\tau\}_{\tau \in \{t-T,t-1\}}, s'_t), P(\{s_\tau\}_{\tau \in \{t-T,t-1\}}, s'_t))$$

↑ *Experiential learning*   ↓ *Decision making*

Space of states and episodic memory ("precedents")

State s_{t-1} (past)	State s_t (present)		State s_{t+1} (future)
Reality ("fact") s_{t-1}	Expectation ("plan") s'_t	Reality ("fact") s_t	Expectation ("plan") s'_{t+1}
Sensations f_{t-1}	Sensations f'_t	Sensations f_t	Sensations f'_{t+1}
Needs y_{t-1}	Needs y'_t	Needs y_t	Needs y'_{t+1}
Actions a_{t-1}	Actions a'_t	Actions a_t	Actions a'_{t+1}

Decisions   *Perceptions*

Actuators   Sensors

Psyche = Operating system

Energy-efficiency *of a*

Optimal decision and expectation of *s*

Predictability *of s*

Expected utility of *s*

Intelligence = Decision making system

Models *s* ("invariants") of states with utilities *U* and probabilities *P* of transitions

$$U(\{s_\tau\}_{\tau \in \{-T,-1\}}, s'_0) = L(\ x \cdot (y_t - y_{t-1}),\ 1-(s'_t - s_t),\ 1-E(a_{t-1}))\qquad s'_t = argmax_s(U(\{s_\tau\}_{\tau \in \{t-T,t-1\}}, s'_t), P(\{s_\tau\}_{\tau \in \{t-T,t-1\}}, s'_t))$$

Satisfaction *of y*

Expected probability of *s*

↑ *Experiential learning*   ↓ *Decision making*

Space of states and episodic memory ("precedents")

| State s_{t-1} (past)	State s_t (present)		State s_{t+1} (future)
Reality ("fact") s_{t-1}	**Expectation ("plan")** s'_t	**Reality ("fact")** s_t	**Expectation ("plan")** s'_{t+1}
Sensations f_{t-1}	Sensations f'_t	Sensations f_t	Sensations f'_{t+1}
Needs y_{t-1}	Needs y'_t	Needs y_t	Needs y'_{t+1}
Actions a_{t-1}	Actions a'_t	Actions a_t	Actions a'_{t+1}

Decisions   *Perceptions*

Actuators   Sensors

$x \cdot y_t$ – "motivation vector"
V. F. Petrenko and A. P. Suprun, "Goal oriented systems, evolution, and the subjective aspect in systemology," Tr. Inst. Sistem. Analiza RAN 62 (1) (2012)

Decision making as operational risk management
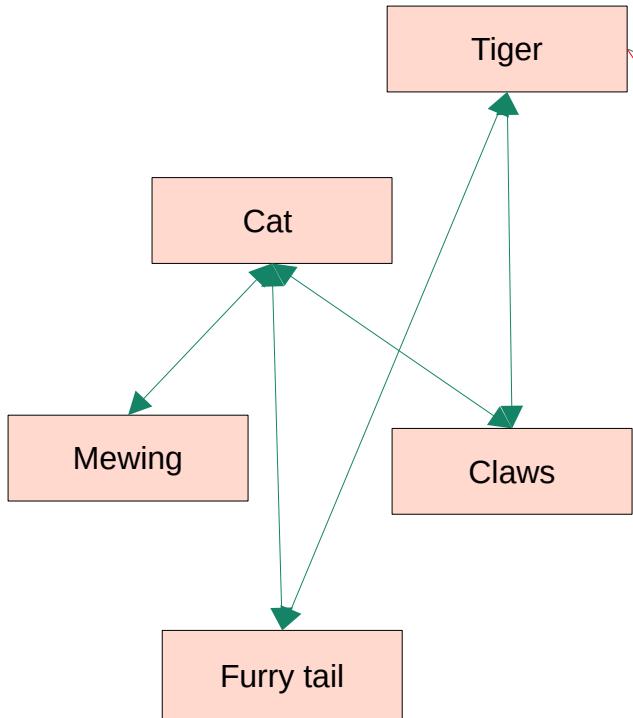
s_t	s'_{t+1}	s'_{t+1}			U	P	$\sum U*P$
		a'	y'	f'			
s_1	s'_2	a'_1	y'_1	...	1.0	0.5	**0.7**
s_1	s'_3	a'_1	y'_2	...	0.4	0.5	
s_1	s'_4	a'_2	y'_3	...	1.0	0.8	**0.8**
s_1	s'_5	a'_2	y'_4	...	0.0	0.2	
s_1	s'_6	a'_3	y'_5	...	0.6	1.0	**0.6**
s_1	s'_7	a'_4	y'_6	...	0.0	1.0	**0.0**

$s'_t = argmax_s(U,P)$
$a'_t = argmax_a(U,P)$

a'_1 a'_2 a'_3 a'_4
s'_2 s'_4 s'_6 s'_7
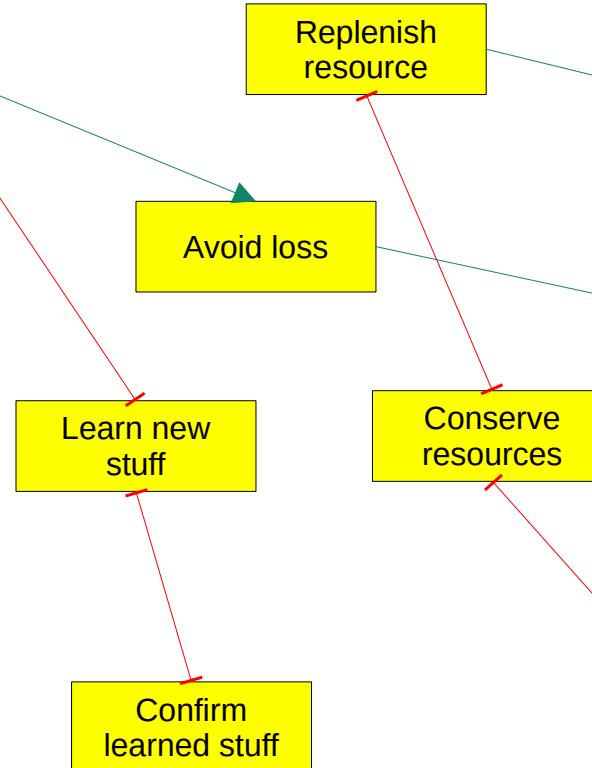
☺ Tversky & Kahneman: most people choose a'_3 и s'_6 ("smaller profit with greater reliability")

# Mutual dependency of state variable ~~subgraphs~~ tensors

Percept feelings: Tiger, Cat, Mewing, Claws, Furry tail

Satisfied needs: Replenish resource, Avoid loss, Learn new stuff, Conserve resources, Confirm learned stuff
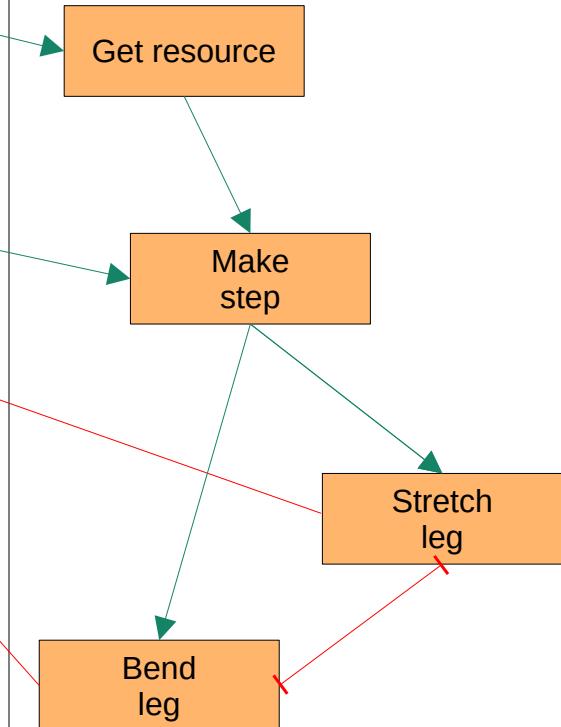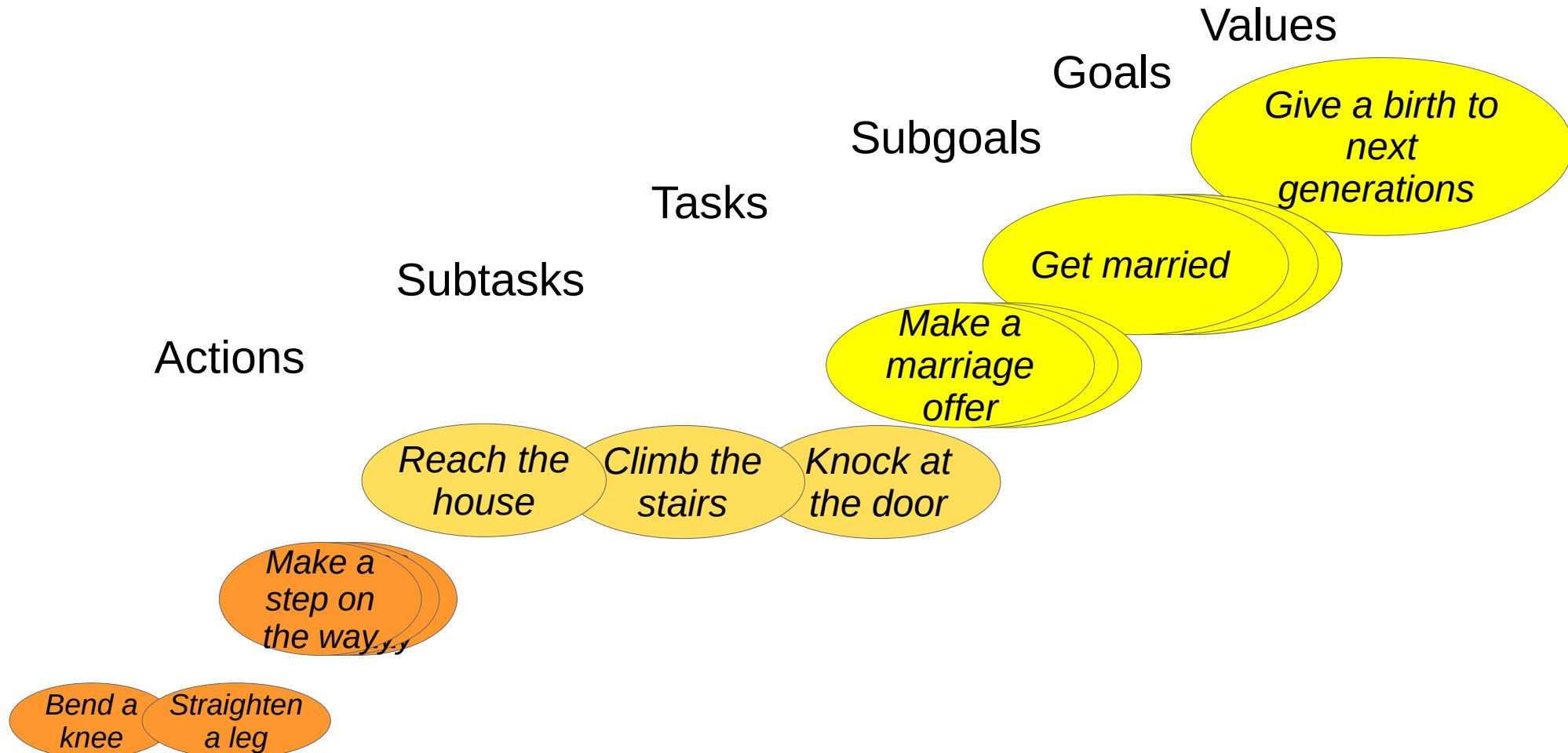
Decided & committed actions: Get resource, Make step, Stretch leg, Bend leg

13
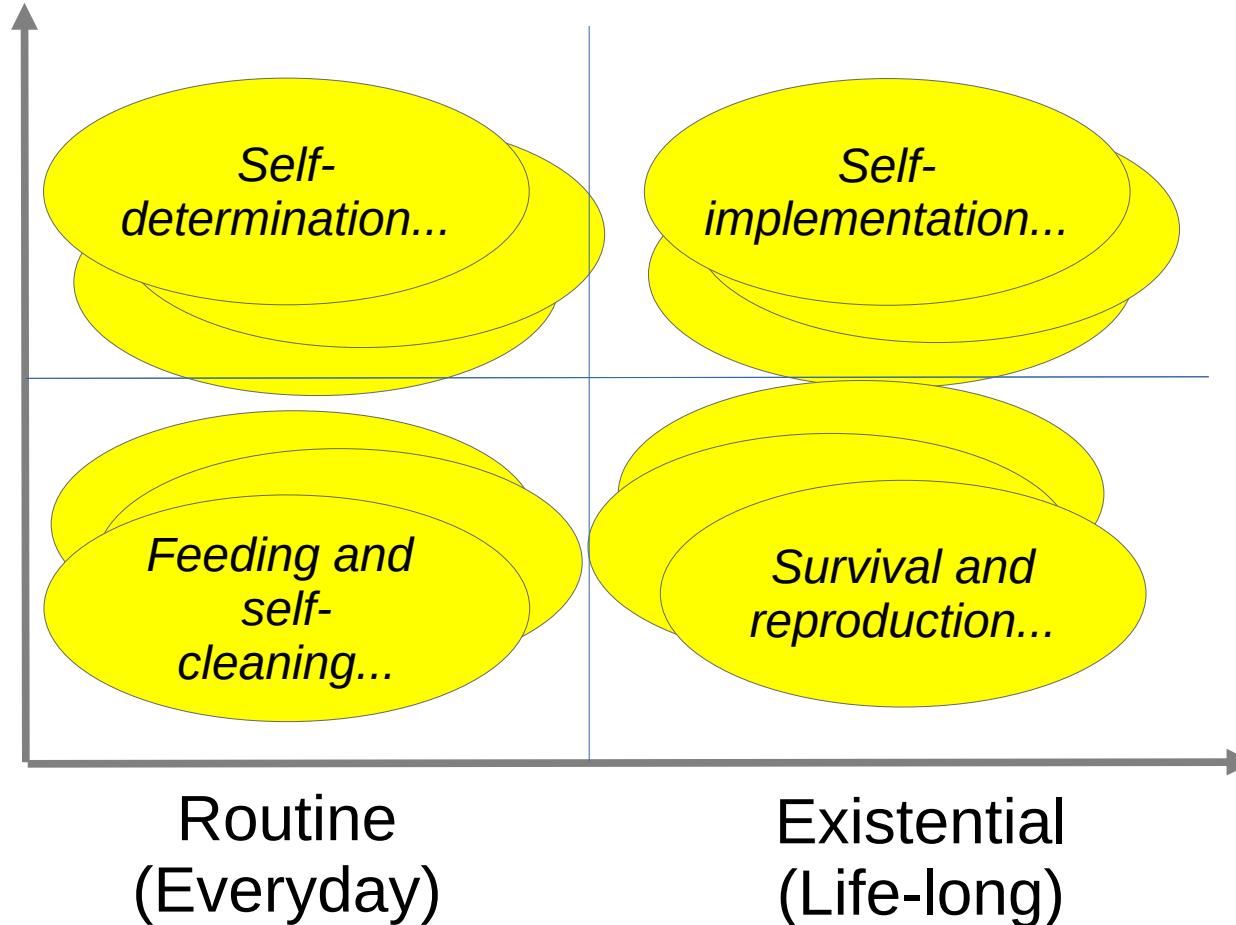
Hierarchy of values/goals/subgoals/tasks/subtasks

Values

Goals

Subgoals

Tasks

Subtasks

Actions

Give a birth to next generations

Get married

Make a marriage offer

Reach the house

Climb the stairs

Knock at the door

Make a step on the way...

Bend a knee

Straighten a leg

14

Space of needs/goals/values

Psychological

Self-determination...

Self-implementation...
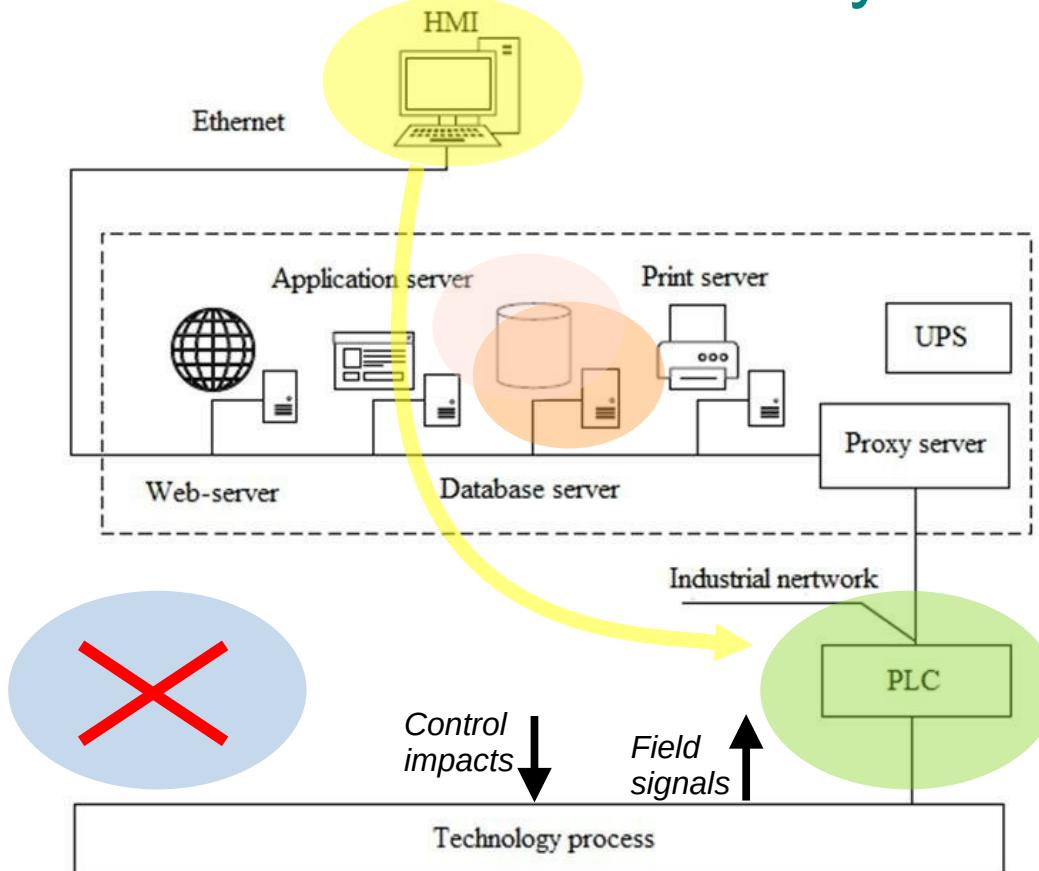
Physical (Physiological)

Feeding and self-cleaning...

Survival and reproduction...

According to the ideas of A. Maslow, any goal can be represented as a combination (superposition) of basic needs

Routine (Everyday)

Existential (Life-long)
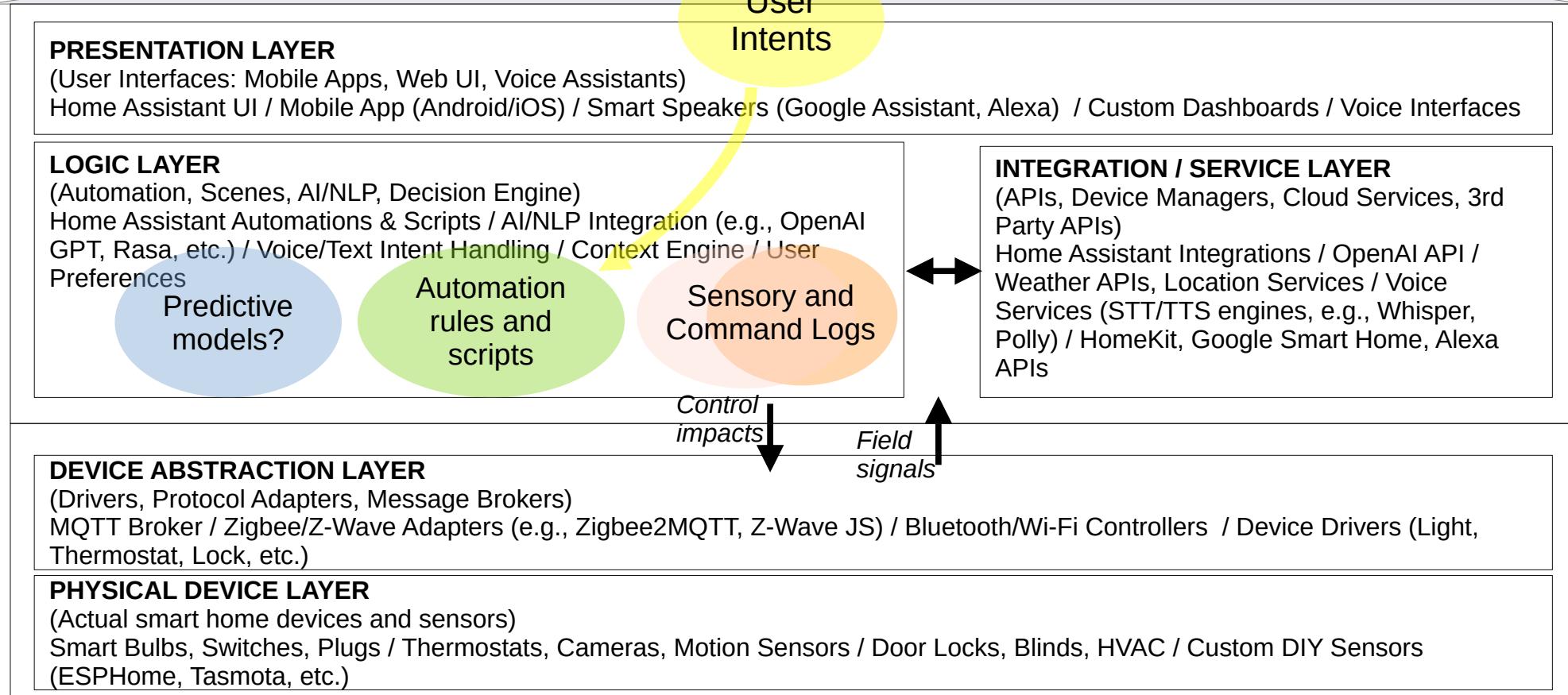
Application cases
A) Automated Process Control Systems (APCS)

Programmable logic

IEC 61131 (synchronous)

IEC 61499 (asynchronous)

https://www.researchgate.net/publication/311662442_Adaptive_Intelligent_Manufacturing_Control_Systems

16

Application cases
B) "Smart Home"

User Intents

PRESENTATION LAYER
(User Interfaces: Mobile Apps, Web UI, Voice Assistants)
Home Assistant UI / Mobile App (Android/iOS) / Smart Speakers (Google Assistant, Alexa) / Custom Dashboards / Voice Interfaces

LOGIC LAYER
(Automation, Scenes, AI/NLP, Decision Engine)
Home Assistant Automations & Scripts / AI/NLP Integration (e.g., OpenAI GPT, Rasa, etc.) / Voice/Text Intent Handling / Context Engine / User Preferences

- Predictive models?
- Automation rules and scripts
- Sensory and Command Logs

INTEGRATION / SERVICE LAYER
(APIs, Device Managers, Cloud Services, 3rd Party APIs)
Home Assistant Integrations / OpenAI API / Weather APIs, Location Services / Voice Services (STT/TTS engines, e.g., Whisper, Polly) / HomeKit, Google Smart Home, Alexa APIs

Control impacts

Field signals

DEVICE ABSTRACTION LAYER
(Drivers, Protocol Adapters, Message Brokers)
MQTT Broker / Zigbee/Z-Wave Adapters (e.g., Zigbee2MQTT, Z-Wave JS) / Bluetooth/Wi-Fi Controllers / Device Drivers (Light, Thermostat, Lock, etc.)

PHYSICAL DEVICE LAYER
(Actual smart home devices and sensors)
Smart Bulbs, Switches, Plugs / Thermostats, Cameras, Motion Sensors / Door Locks, Blinds, HVAC / Custom DIY Sensors (ESPHome, Tasmota, etc.)

Implementation options
1) Reinforcement learning based on "actor-critic" model

https://medium.com/intro-to-artificial-intelligence/the-actor-critic-reinforcement-learning-algorithm-c8095a655c14
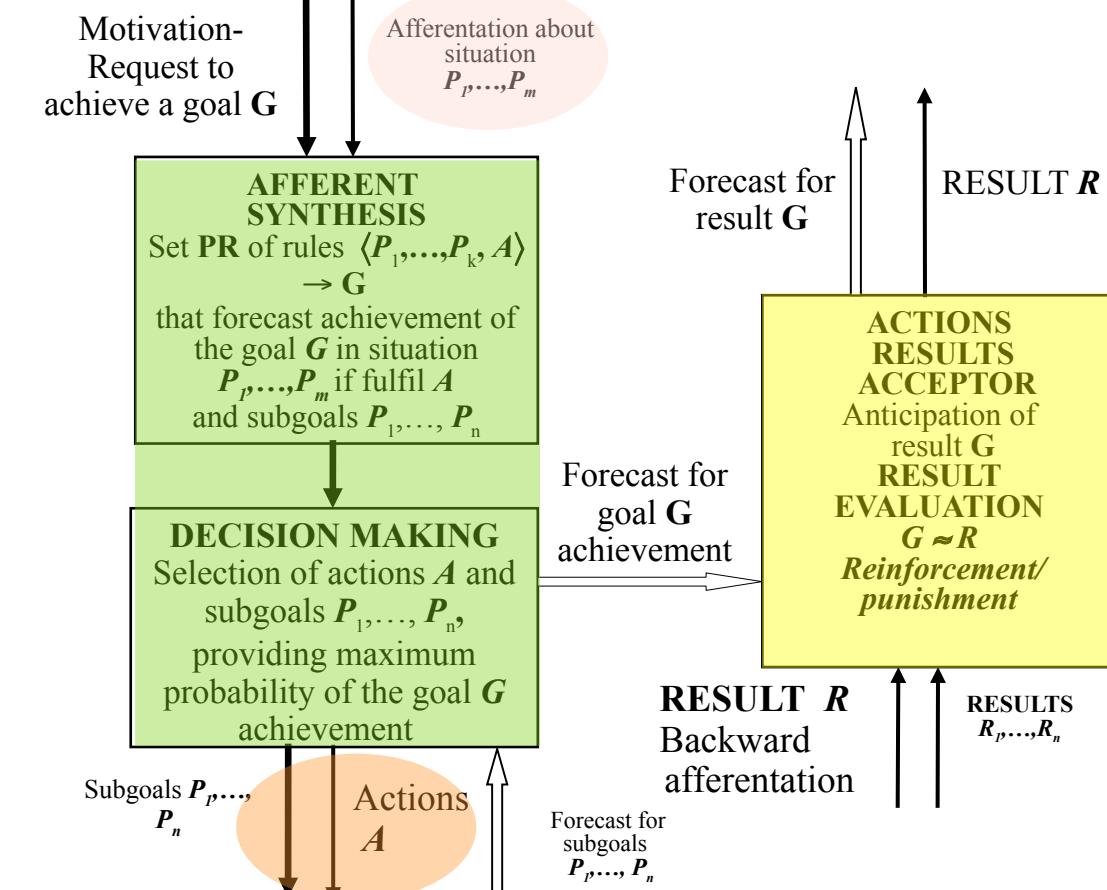
Implementation options

2) Cognitive architecture based on LLM and (active) GraphRAG

Implementation options

3) Cognitive architecture based on probabilistic logic

Motivation-Request to achieve a goal **G**

Afferentation about situation $P_1, \dots, P_m$

AFFERENT SYNTHESIS
Set **PR** of rules $\langle P_1, \dots, P_k, A \rangle \rightarrow G$ that forecast achievement of the goal G in situation $P_1, \dots, P_m$ if fulfil A and subgoals $P_1, \dots, P_n$

DECISION MAKING
Selection of actions A and subgoals $P_1, \dots, P_n$, providing maximum probability of the goal G achievement

Subgoals $P_1, \dots, P_n$

Actions A

Forecast for subgoals $P_1, \dots, P_n$

Forecast for goal **G** achievement

Forecast for result **G**

RESULT **R**

ACTIONS RESULTS ACCEPTOR
Anticipation of result **G**
RESULT EVALUATION
$G \approx R$
Reinforcement/ punishment

RESULT R
Backward afferentation

RESULTS $R_1, \dots, R_n$

$$Prob(G \mid P_1, \dots, P_k, P_1, \dots, P_n, A) = Prob(rule) \cdot Prob(P_1) \cdot \dots \cdot Prob(P_n)$$

Evgenii Vityaev, Alexander Demin: Adaptive Control of Modular Robots // Conference Paper in Advances in Intelligent Systems and Computing, Conference: First International Early Research Career Enhancement School on Biologically Inspired Cognitive Architectures, Springer, August 2018

Evgenii E. Vityaev: Purposefulness as a Principle of Brain Activity // Anticipation: Learning from the Past, (ed.) M. Nadin. Cognitive Systems Monographs, V.25, Chapter No.: 13. Springer, 2015, pp. 231-254.

Витяев Е.Е. Логика работы мозга. Подходы к моделированию мышления. (сборник под ред. д.ф.-м.н. В.Г. Редько). УРСС Эдиториал, Москва, 2014г., стр. 120-153.
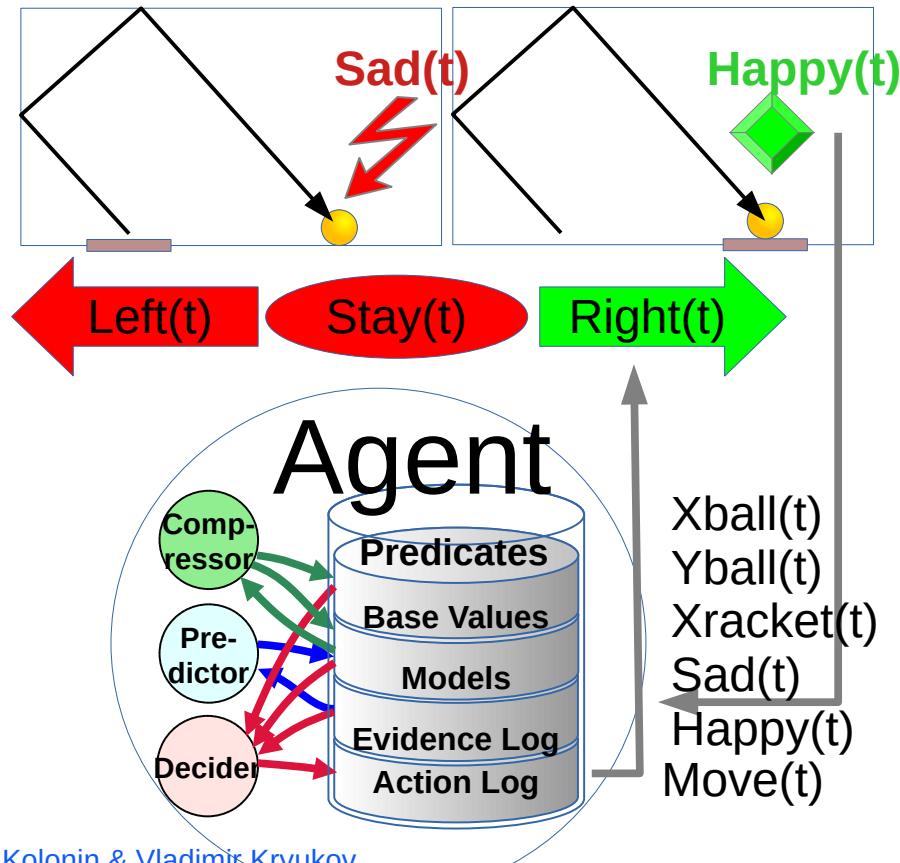
Implementation options
4) Cognitive architecture of value-based experiential learning
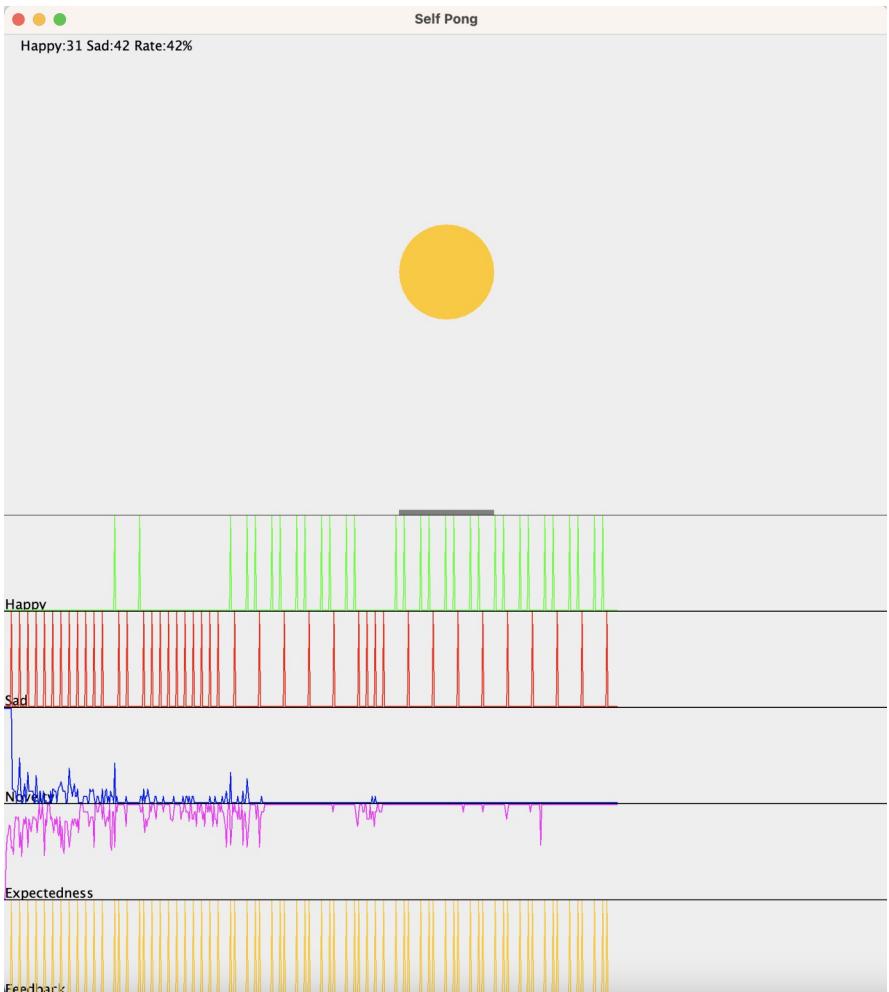
Kolonin, A. (2022). Neuro-Symbolic Architecture for Experiential Learning in Discrete and Functional Environments. In: Goertzel, B., Iklé, M., Potapov, A. (eds) Artificial General Intelligence. AGI 2021. Lecture Notes in Computer Science(), vol 13154. Springer, Cham. https://doi.org/10.1007/978-3-030-93758-4_12

Cognitive architecture of value-based experiential learning

Sad(t)

Happy(t)

Left(t)   Stay(t)   Right(t)

Agent

Comp-ressor

Pre-dictor

Decider

Predicates
Base Values
Models
Evidence Log
Action Log

Xball(t)
Yball(t)
Xracket(t)
Sad(t)
Happy(t)
Move(t)

Anton Kolonin & Vladimir Kryukov,
Computational Concept of the Psyche,
Neuroinformatics-2025
https://arxiv.org/pdf/2509.07009

Self Pong

Happy:31 Sad:42 Rate:42%

Happy

Sad

Novelty

Expectedness

Feedback

Thank you for attention! Questions?

Anton Kolonin

akolonin@aigents.com

Telegram/GitHub: akolonin

Workshop recording
on the subject

Anton Kolonin & Vladimir Kryukov,
Computational Concept of the
Psyche, Neuroinformatics-2025