On Cognitive Architectures for Interpretable Strong/General AI

Anton Kolonin
akolonin@aigents.com
Facebook: akolonin
Telegram: akolonin

AIGENTS
https://aigents.com

AGI
IN RUSSIAN
https://facebook.com/groups/agirussia
https://t.me/agirussia

Novosibirsk State University
*THE REAL SCIENCE

SingularityNET
https://singularitynet.io

Definition of General Intelligence

Importance of Interpretability

Consciousness – Ontological Modeling

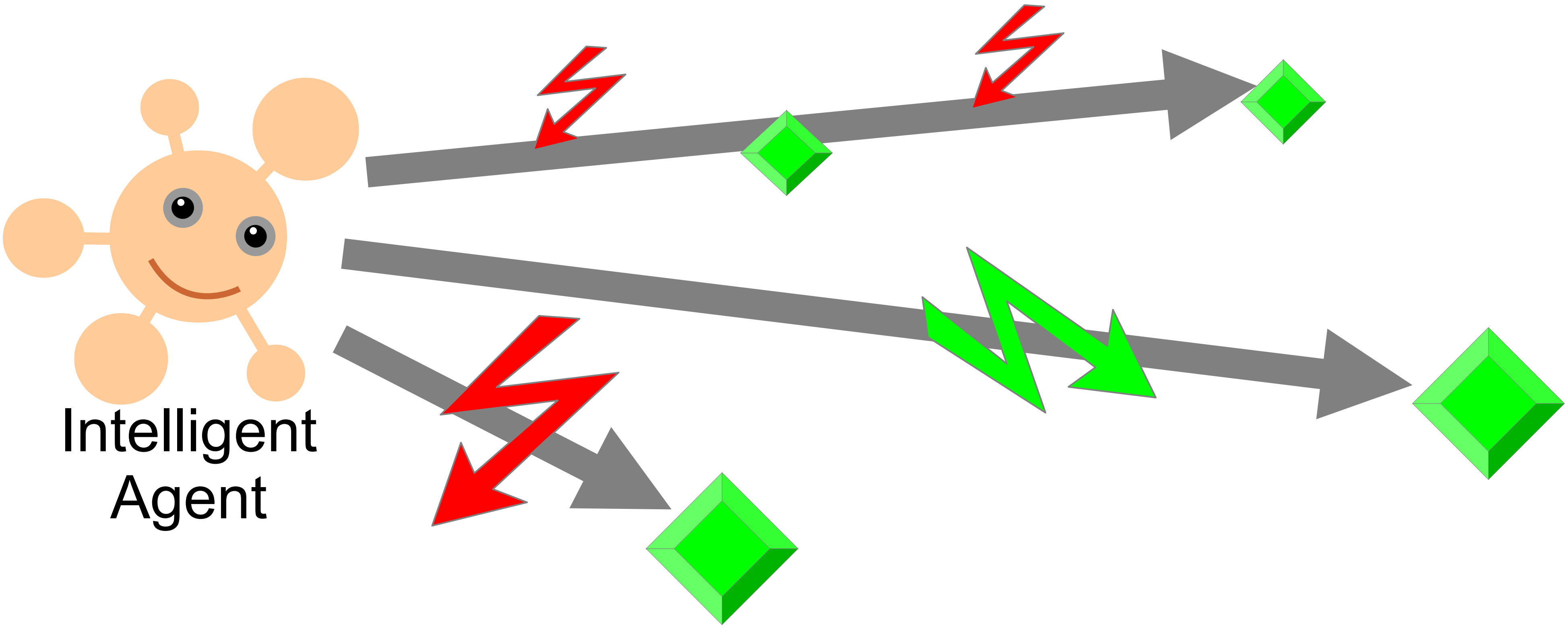
Neuro-Symbolic Architectures

Simulation Results and Conclusions

General Intelligence:

Reaching complex goals in different complex environments, using limited resources and minimizing risks

(Ben Goertzel + Pei Wang + Shane Legg + Marcus Hutter)

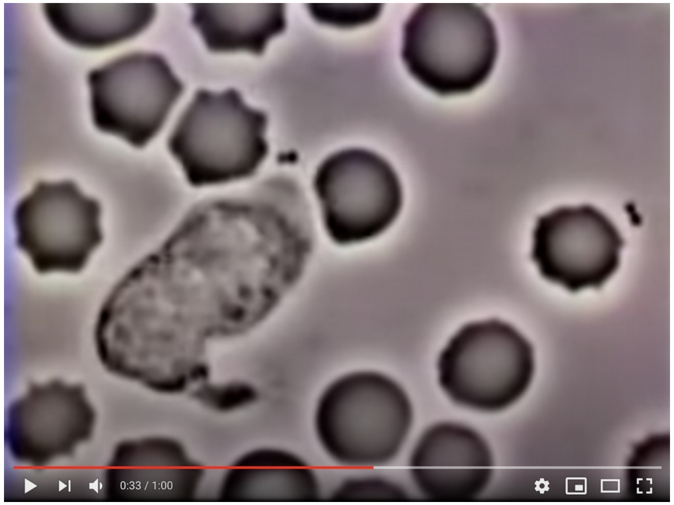


Minimally viable natural system capable to satisfy the requirement?

Complex nervous system

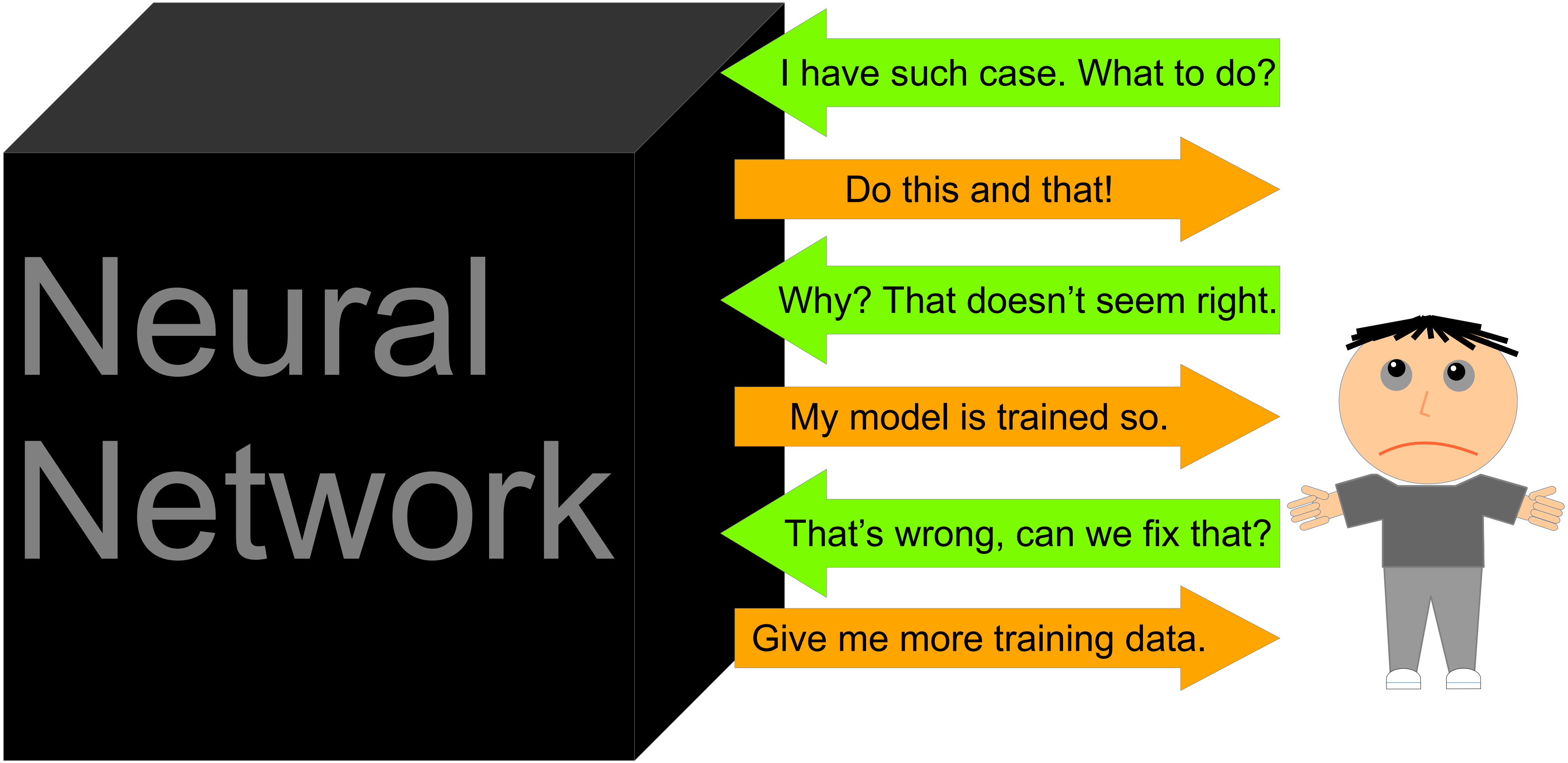
Simple nervous system

Single cell organism



Interpretability vs. Explainability

Can we trust to what we do not understand?
Can we know what has been have learned?
Can we tell them what we need exactly?



Neural Network

I have such case. What to do?

Do this and that!

Why? That doesn’t seem right.

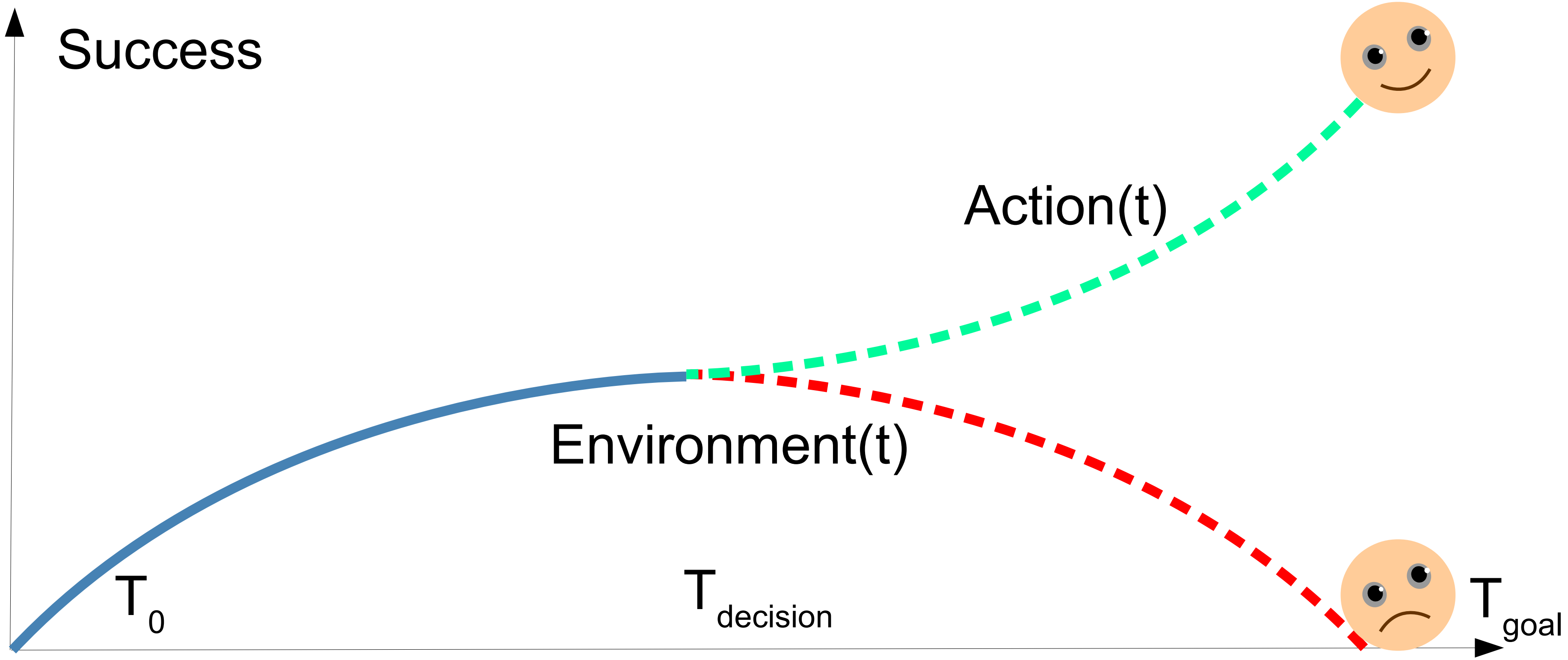
My model is trained so.

That’s wrong, can we fix that?

Give me more training data.

Consciousness:

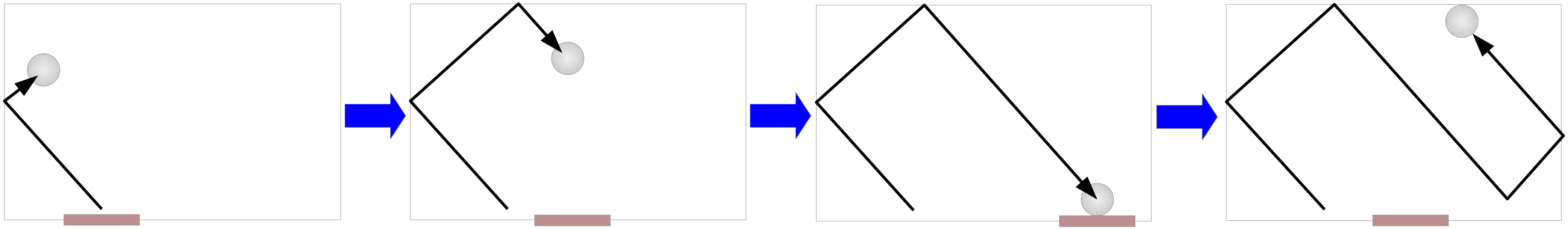
Ability to build models of the environment based on the past to predict the future scenarios and act “consciously” towards the desired ones



Copyright © 2021 Anton Kolonin, Aigents®

Acting consciously:

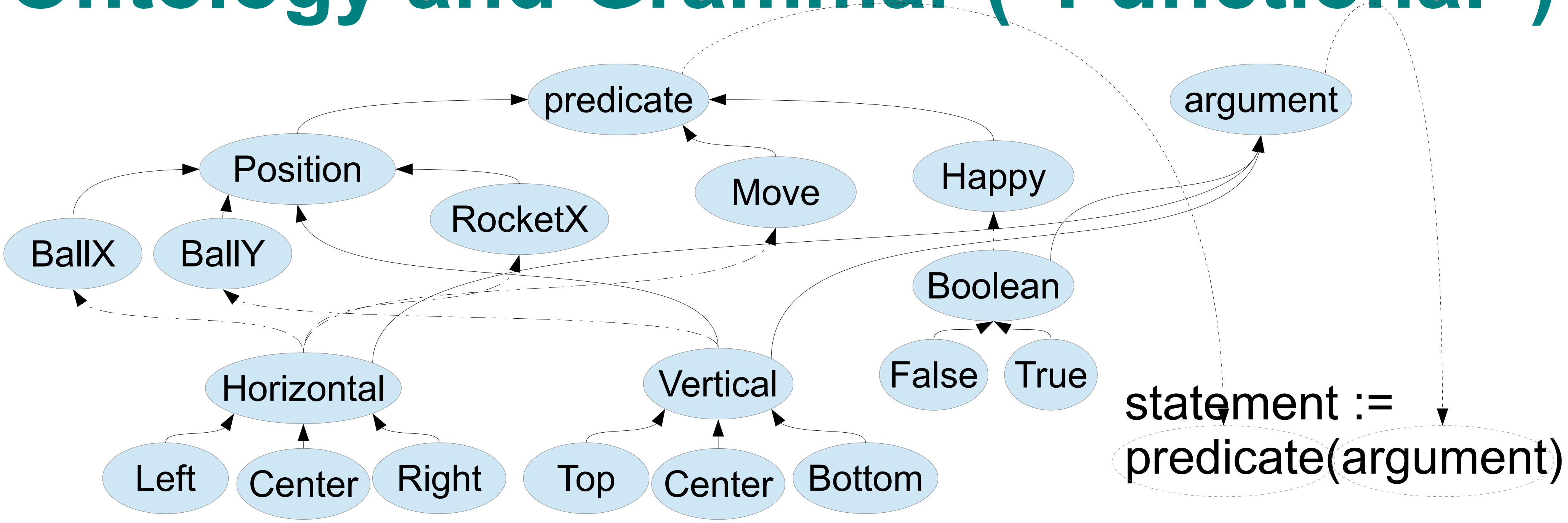
Agent being able to execute the sequence of behavioral acts to itself by means of a language (system of predicates within an ontology)



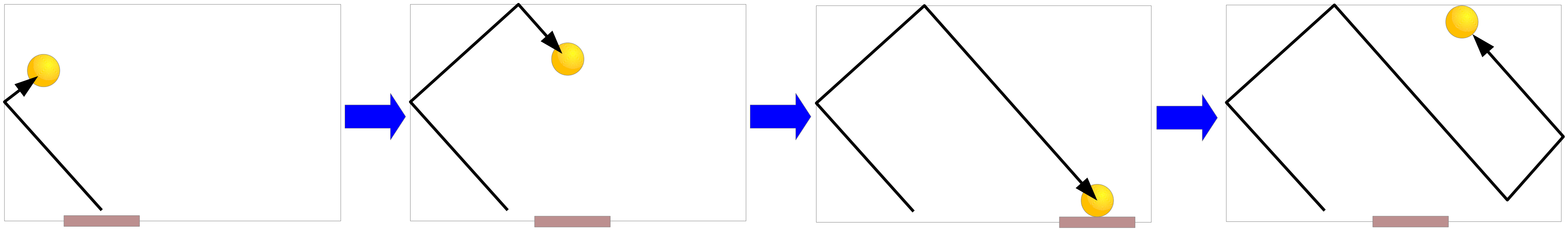
BallY(Top)	BallY(Top)	BallY(Bottom)	BallY(Bottom)
BallX(Left)	BallX(Center)	BallX(Right)	BallX(Right)
RocketX(Left)	RocketX(Center)	RocketX(Right)	RocketX(Right)
Happy(False)	Happy(False)	Happy(False)	**Happy(True)**
=>	=>	=>	=>
Move(Right)	Move(Right)	Move(Left)	Move(Left)

https://www.youtube.com/watch?v=2LPLhJKh95g
https://github.com/aigents/aigents-java/tree/master/src/main/java/net/webstructor/agi

Ontology and Grammar (“Functional”)



statement :=
predicate(argument)



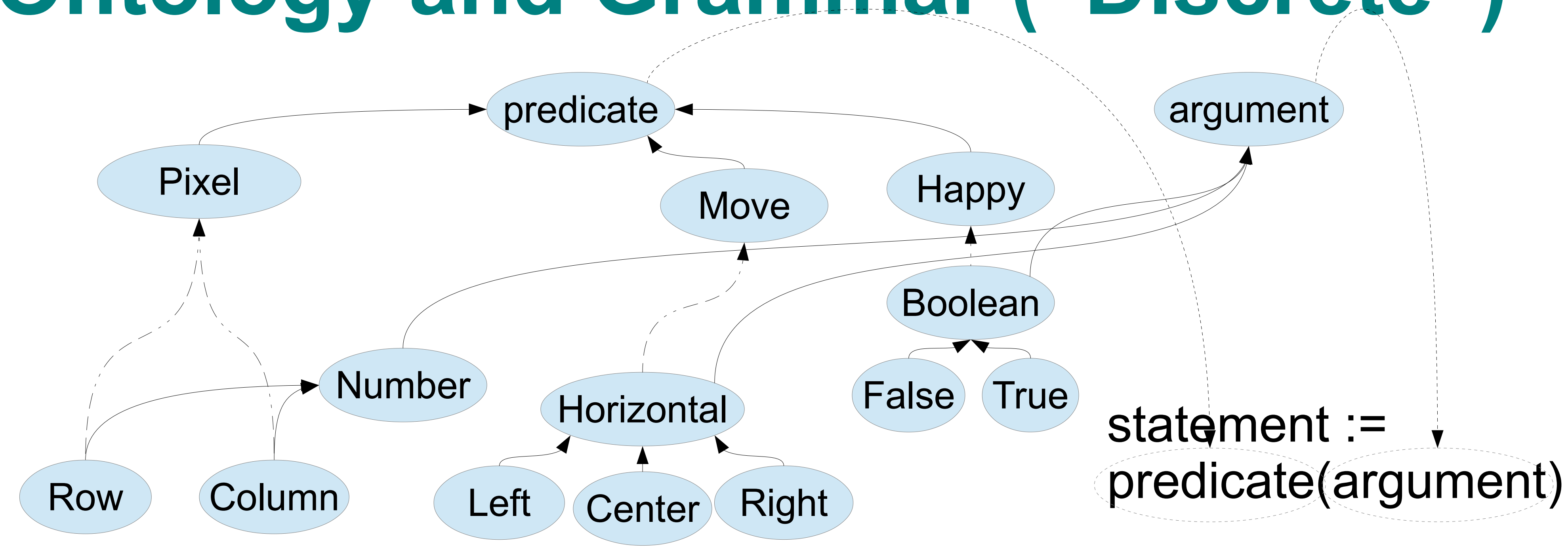
BallY(Top)
BallX(Left)
RocketX(Left)
Happy(False)
=>
Move(Right)

BallY(Top)
BallX(Center)
RocketX(Center)
Happy(False)
=>
Move(Right)

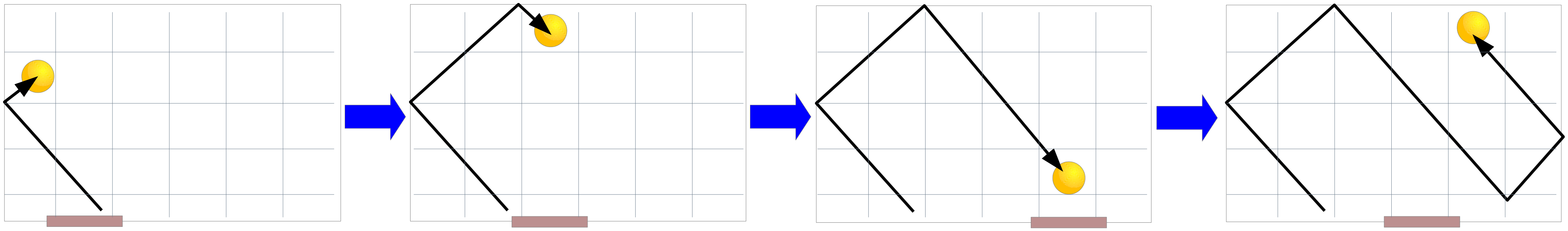
BallY(Bottom)
BallX(Right)
RocketX(Right)
Happy(False)
=>
Move(Left)

BallY(Bottom)
BallX(Right)
RocketX(Right)
Happy(True)
=>
Move(Left)

Ontology and Grammar (“Discrete”)



statement :=
predicate(argument)



Pixel(1,0)
Pixel(4,1)
Happy(False)
=>
Move(Right)

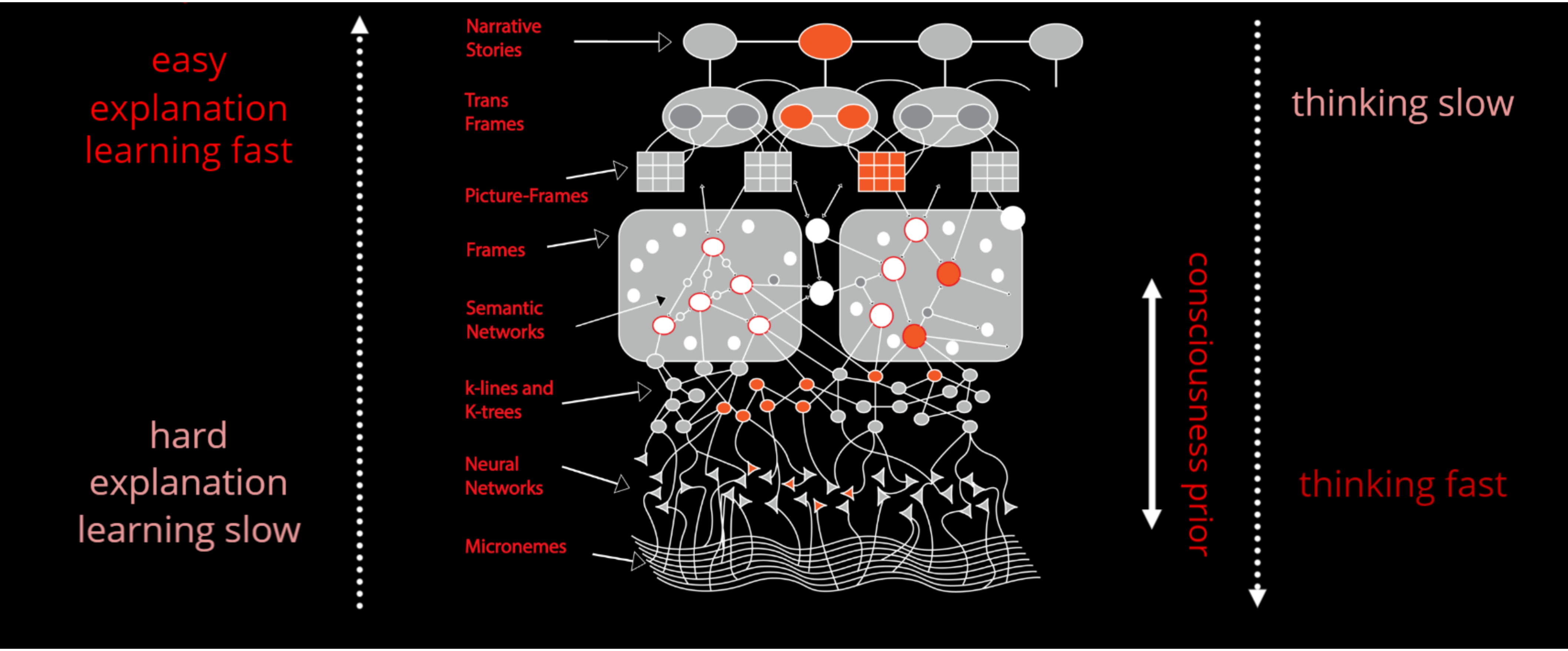
Pixel(0,2)
Pixel(4,2)
Happy(False)
=>
Move(Right)

Pixel(3,4)
Pixel(4,4)
Happy(False)
=>
Move(Left)

Pixel(0,4)
Pixel(4,3)
Happy(True)
=>
Move(Left)

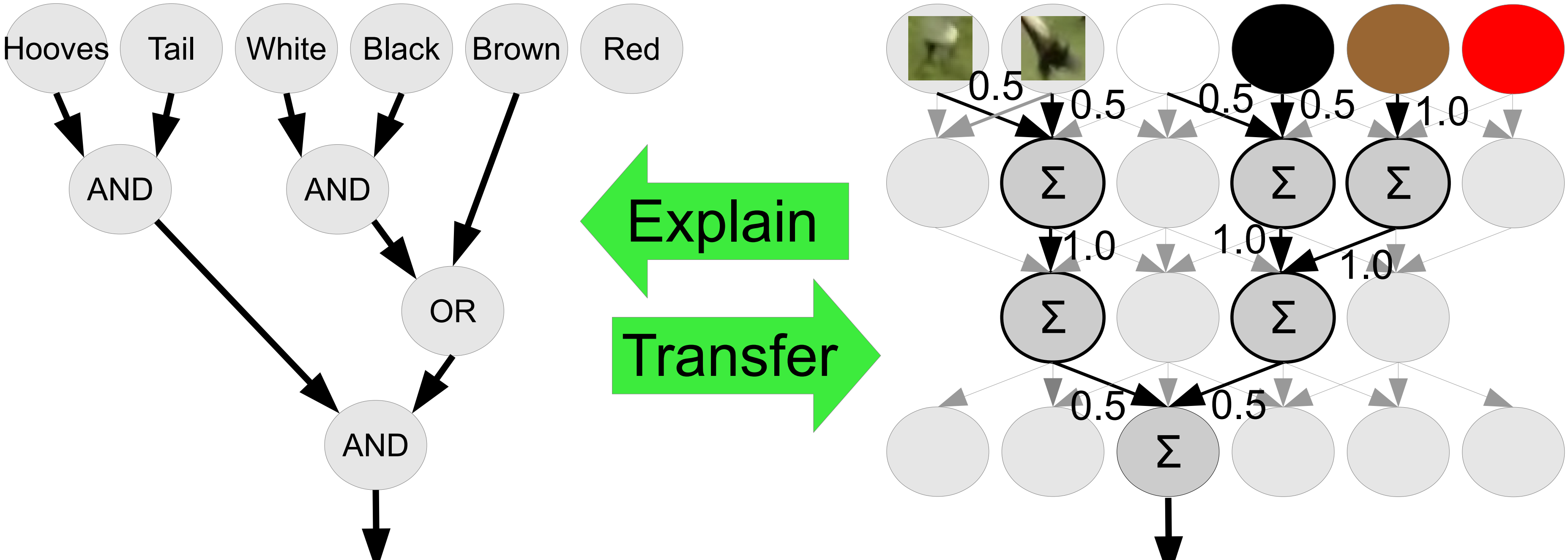
Hybrid Neuro-Symbolic Cognitive Architectures
“Vertical” Neuro-Symbolic Integration

Society of Mind – Marvin Minsky
Thinking, Fast and Slow – Daniel Kahneman



https://towardsdatascience.com/explainable-ai-vs-explaining-ai-part-1-d39ea5053347

Bridging the Symbolic-Subsymbolic gap for “explainable AI” and “transfer learning” - “Horizontal” Neuro-Symbolic Integration



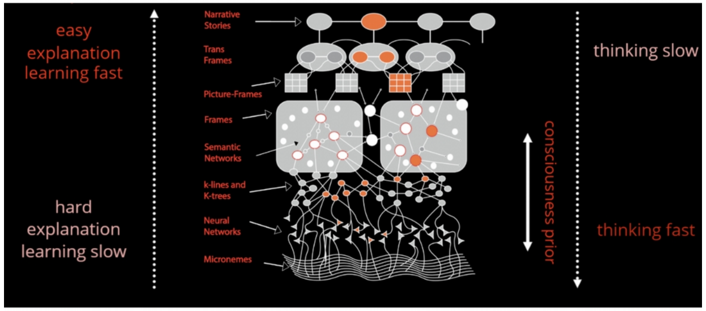
(Hooves AND Tail) AND
((White and Black) OR Brown)

=> Horse

Imaginable AGI Architectures

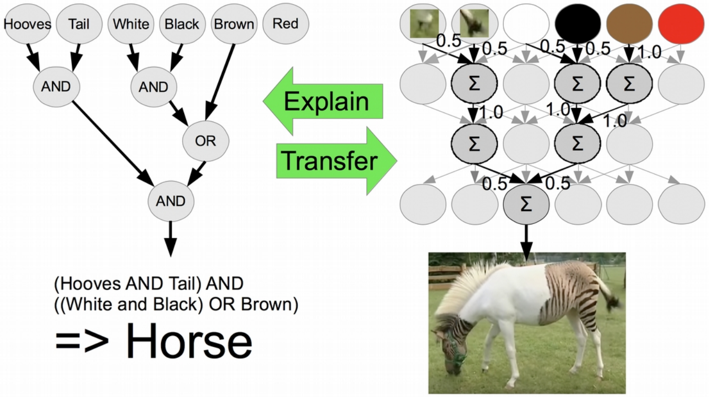
“Vertical” Neuro-Symbolic Integration

Society of Mind – Marvin Minsky
Thinking, Fast and Slow – Daniel Kahneman



https://towardsdatascience.com/explainable-ai-vs-explaining-ai-part-1-d39ea5053347

“Horizontal” Neuro-Symbolic Integration



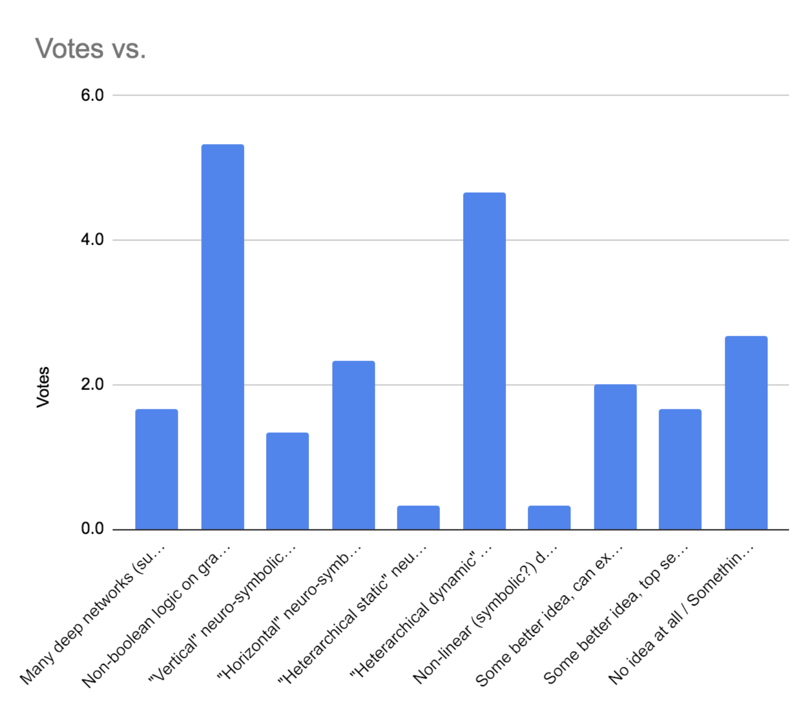
(Hooves AND Tail) AND
((White and Black) OR Brown)
=> Horse

	Votes	AGI	AGIRussia (FB)	AGIRussia (TG)
Many deep networks (subsymbolic)	1.7	1	1	3
Non-boolean logic on graphs/predicates (symbolic)*	5.3	5	9	2
"Vertical" neuro-symbolic integration	1.3	2	1	1
"Horizontal" neuro-symbolic integration	2.3	2	2	3
"Heterarchical static" neuro-symbolic integration	0.3	0	0	1
"Heterarchical dynamic" neuro-symbolic integration**	4.7	7	4	3
Non-linear (symbolic?) dynamic (R.Freeman)	0.3	1	0	0
Some better idea, can exlplain	2.0	0	1	5
Some better idea, top secret	1.7	2	1	2
No idea at all / Something unimaginable yet	2.7	4	0	4

* Evidence-based reasoning (M.Ryabchevsky)

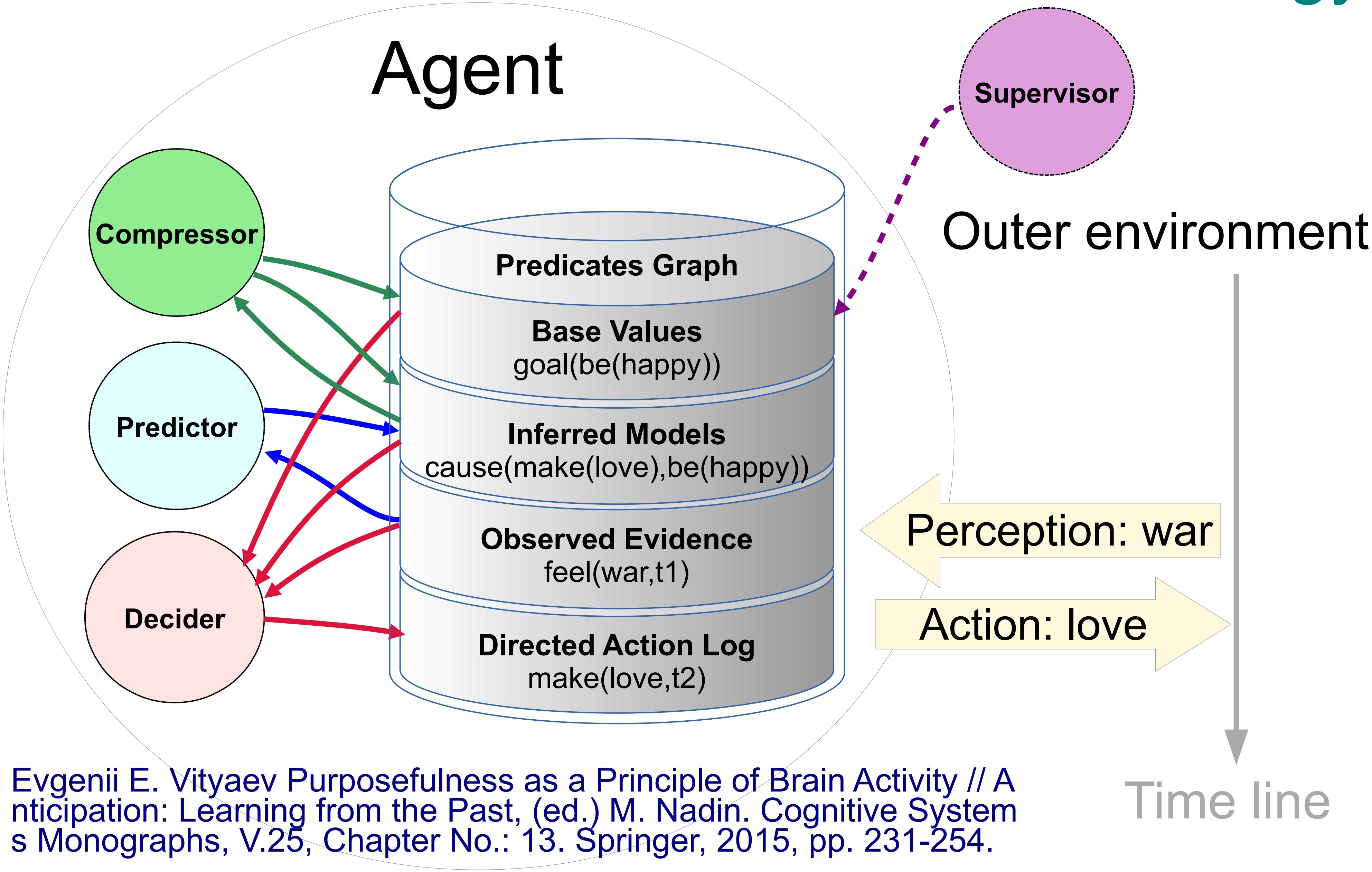
** "Building Minds with Patterns" (M.Miller)

** Architure-agnostic (A.Kabanov)



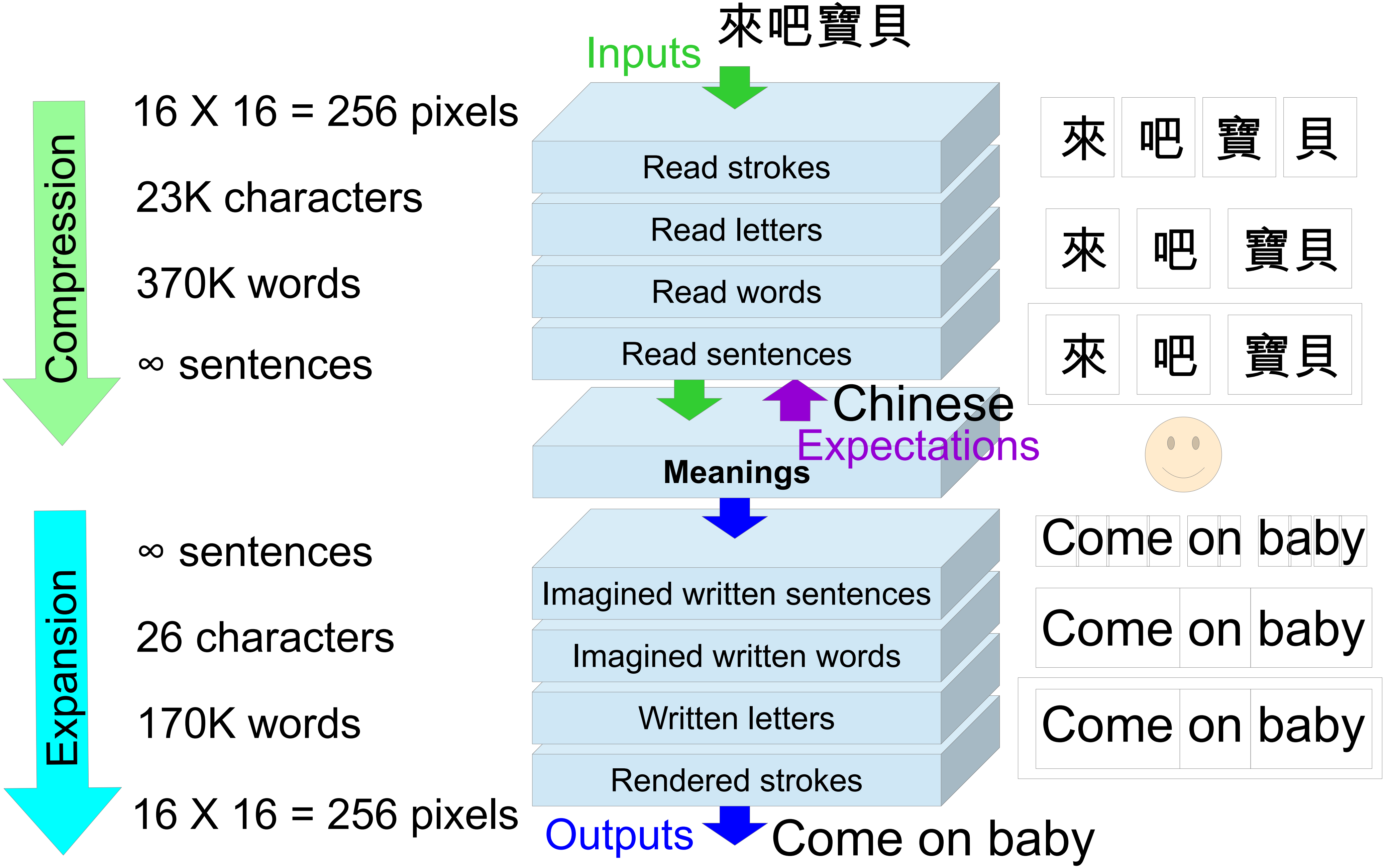
https://docs.google.com/spreadsheets/d/1Ilm3hu9aewpQc-Mjl8xChjkKXr21gnh0aQ74EnhygX4/

An Agent of AGI Cognitive Architecture based on TFS and Environmental Ontology

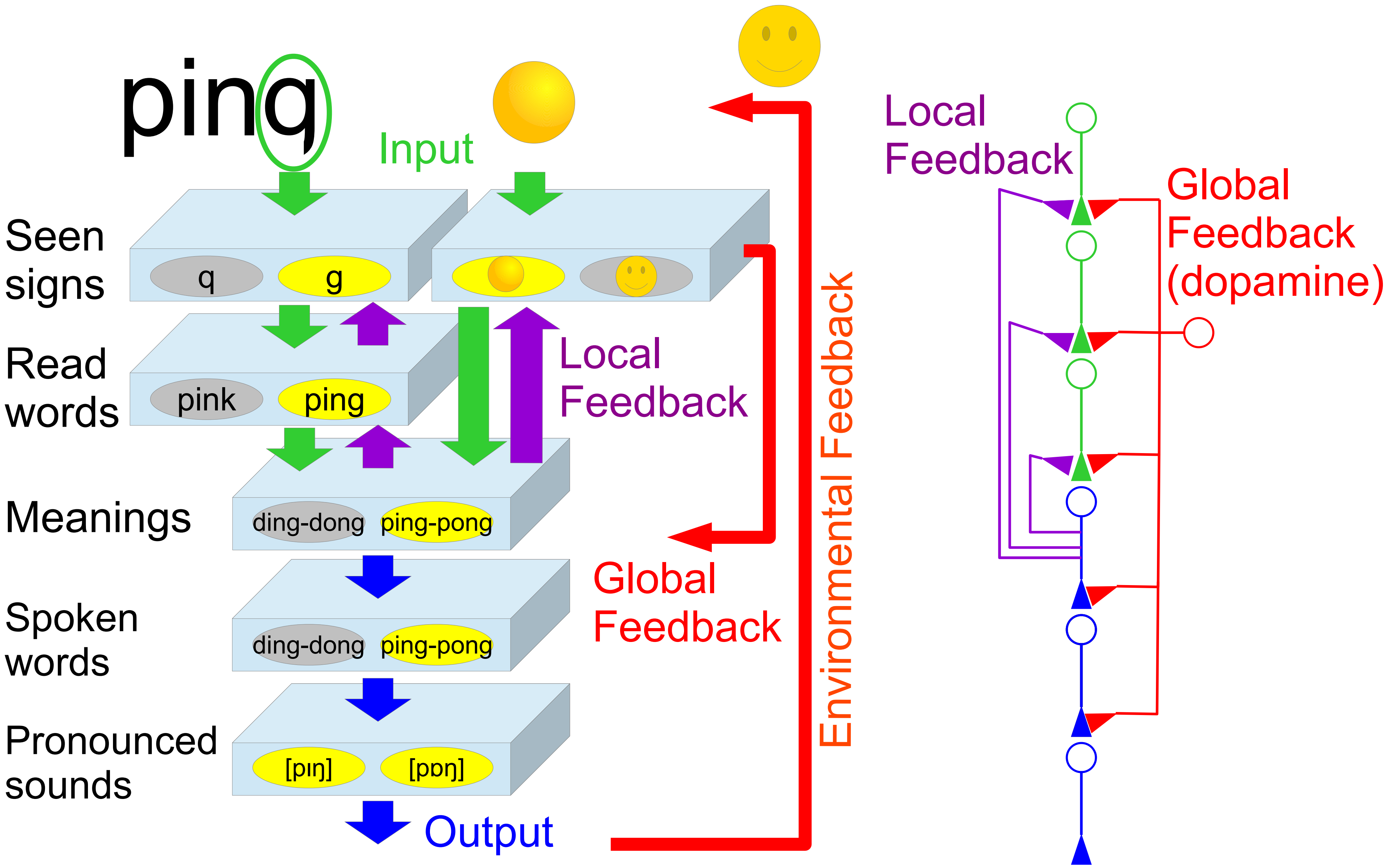


Evgenii E. Vityaev Purposefulness as a Principle of Brain Activity // Anticipation: Learning from the Past, (ed.) M. Nadin. Cognitive Systems Monographs, V.25, Chapter No.: 13. Springer, 2015, pp. 231-254.

Architecture: Multi-layer

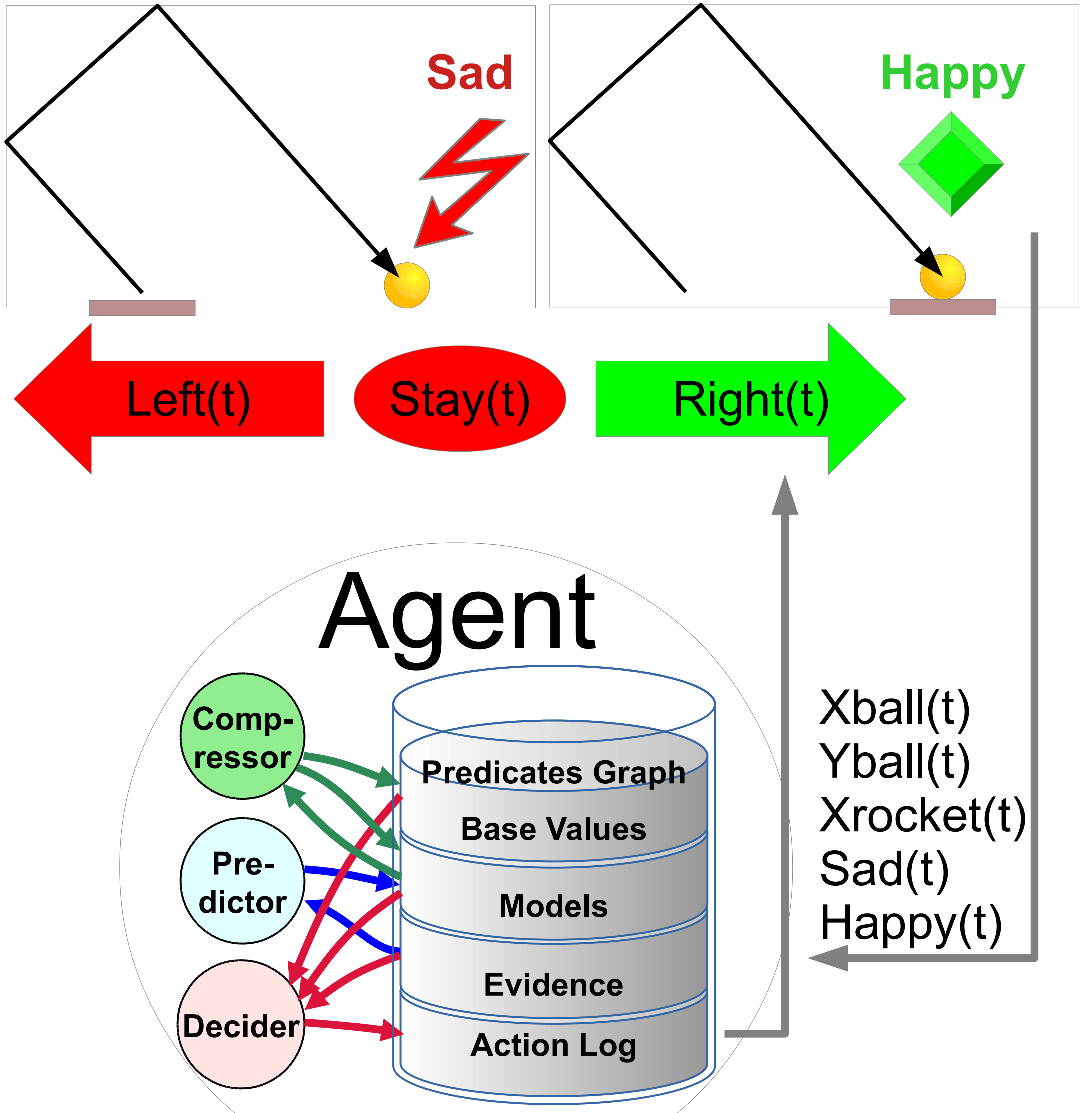


Architecture: Local/Global Feedback



http://www.acad.bg/ebook/ml/Society%20of%20Mind.pdf (*"Global and Local Reward"*)

An Agent of AGI Cognitive Architecture learning single-player “ping-pong” game



Sad

Happy

Left(t)

Stay(t)

Right(t)

Agent

Comp-ressor

Pre-dictor

Decider

Predicates Graph

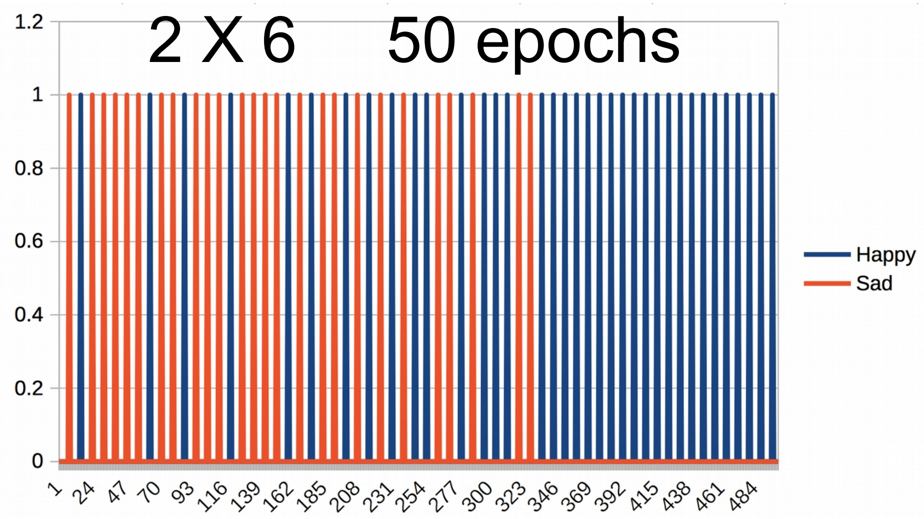
Base Values

Models

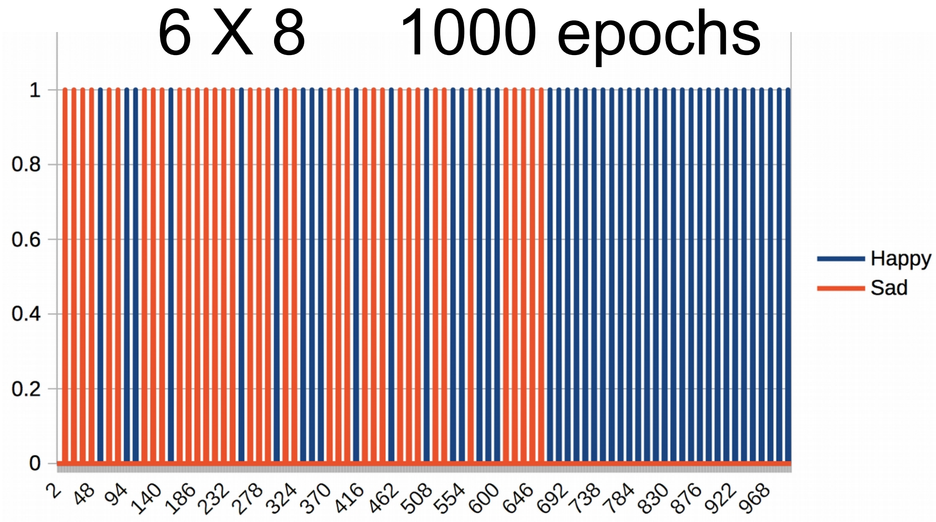
Evidence

Action Log

Xball(t)
Yball(t)
Xrocket(t)
Sad(t)
Happy(t)



2 X 6 50 epochs

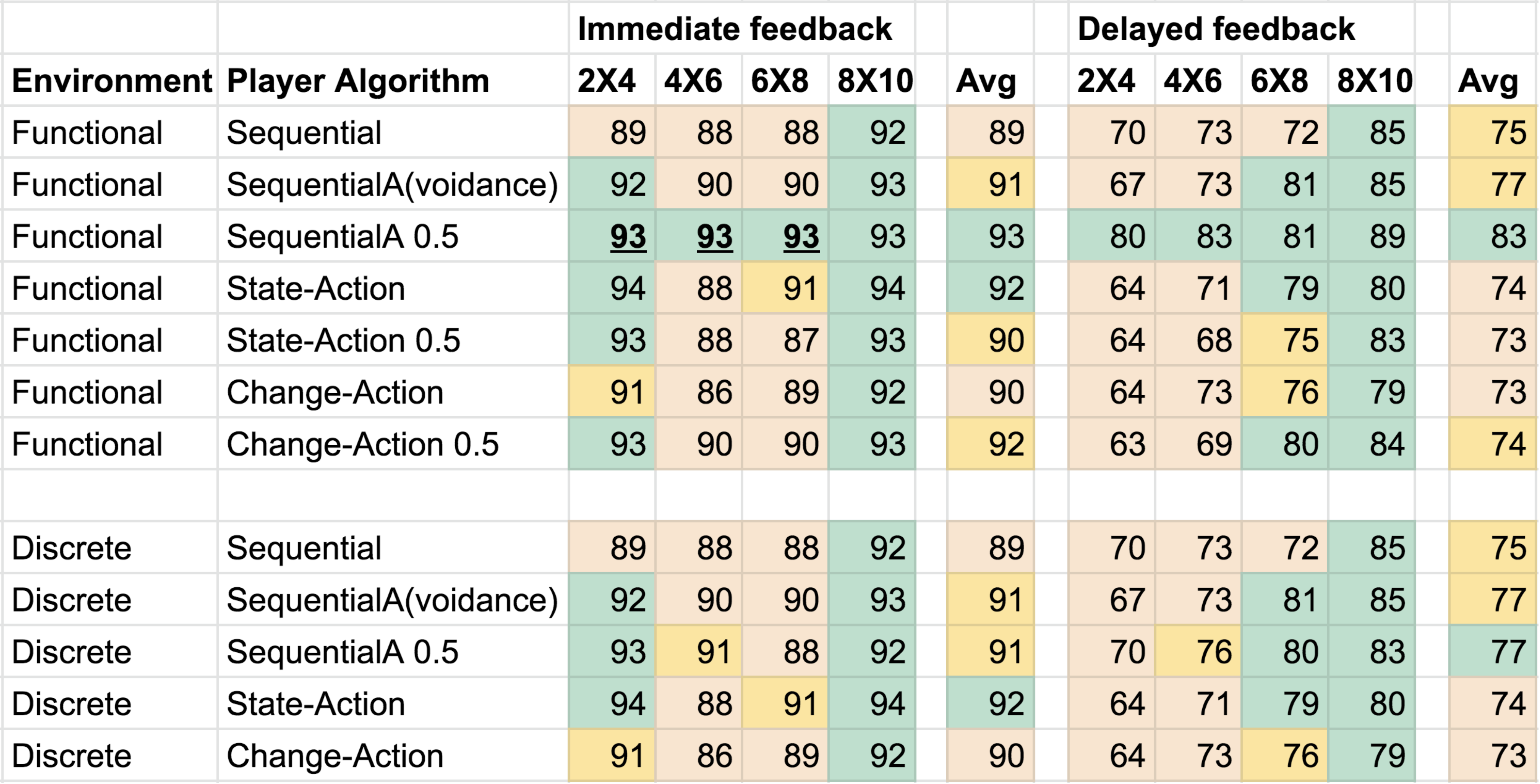


6 X 8 1000 epochs

https://arxiv.org/abs/1807.02072
https://github.com/aigents/aigents-java/blob/master/src/main/java/net/webstructor/util/AgiTester.java

Learning single-player “ping-pong” game

with global feedback for successive behaviors



		Immediate feedback							Delayed feedback					
Environment	**Player Algorithm**	**2X4**	**4X6**	**6X8**	**8X10**		**Avg**		**2X4**	**4X6**	**6X8**	**8X10**		**Avg**
Functional	Sequential	89	88	88	92		89		70	73	72	85		75
Functional	SequentialA(voidance)	92	90	90	93		91		67	73	81	85		77
Functional	SequentialA 0.5	**93**	**93**	**93**	93		93		80	83	81	89		83
Functional	State-Action	94	88	91	94		92		64	71	79	80		74
Functional	State-Action 0.5	93	88	87	93		90		64	68	75	83		73
Functional	Change-Action	91	86	89	92		90		64	73	76	79		73
Functional	Change-Action 0.5	93	90	90	93		92		63	69	80	84		74
Discrete	Sequential	89	88	88	92		89		70	73	72	85		75
Discrete	SequentialA(voidance)	92	90	90	93		91		67	73	81	85		77
Discrete	SequentialA 0.5	93	91	88	92		91		70	76	80	83		77
Discrete	State-Action	94	88	91	94		92		64	71	79	80		74
Discrete	Change-Action	91	86	89	92		90		64	73	76	79		73

https://www.youtube.com/watch?v=2LPLhJKh95g
https://github.com/aigents/aigents-java/tree/master/src/main/java/net/webstructor/agi

Global feedback for successive behaviors - brief preliminary conclusions

1) Both Functional and Discrete representations of the environment are close to be **equivalent** from **accuracy** (learning speed) perspective

2) **Functional representation** is much better from the **run-time performance** (response time and energy saving) perspective

3) Both **avoidance of negative feedback and fuzzy matching** of experiences help are **improving accuracy** and learning speed

4) **Delayed reward decreases accuracy** to extent of ~10-15%

5) Replacing explicit memories of successive behaviors with **global feedback on combinations of state-action and change-action** contexts: a) **increases performance** dramatically, b) **decreases accuracy** a bit.

6) **Negative "global feedback"** makes accuracy significantly **worse**, learning may get impossible in some cases

https://www.youtube.com/watch?v=2LPLhJKh95g
https://github.com/aigents/aigents-java/tree/master/src/main/java/net/webstructor/agi

Thank you and welcome!

Anton Kolonin
akolonin@aigents.com
Facebook: akolonin
Telegram: akolonin

AIGENTS
https://aigents.com

AGI IN RUSSIAN
https://facebook.com/groups/agirussia
https://t.me/agirussia

Novosibirsk State University
*THE REAL SCIENCE

SingularityNET
https://singularitynet.io