Tests and metrics for believable character reasoning inspired by a cognitive architecture

Alexei Samsonovich
(1) "BICA Lab", INSTITUTE of CYBER INTELLIGENCE SYSTEMS (ICIS)
(2) National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
(3) Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA, USA

IJCAI-16 GRW – NYC, July 9, 2016

Opportunities offered by ICIS:

- *Your own international laboratory within NRNU MEPhI.* You can stay in your country and remotely direct or co-direct research, conducted in a newly created laboratory in NRNU MEPhI. Total funds for the lab can be around 10M Rubles per year, and your personal salary can be up to 40-50% of this amount, i.e., approximately $70-$80K USD per year.

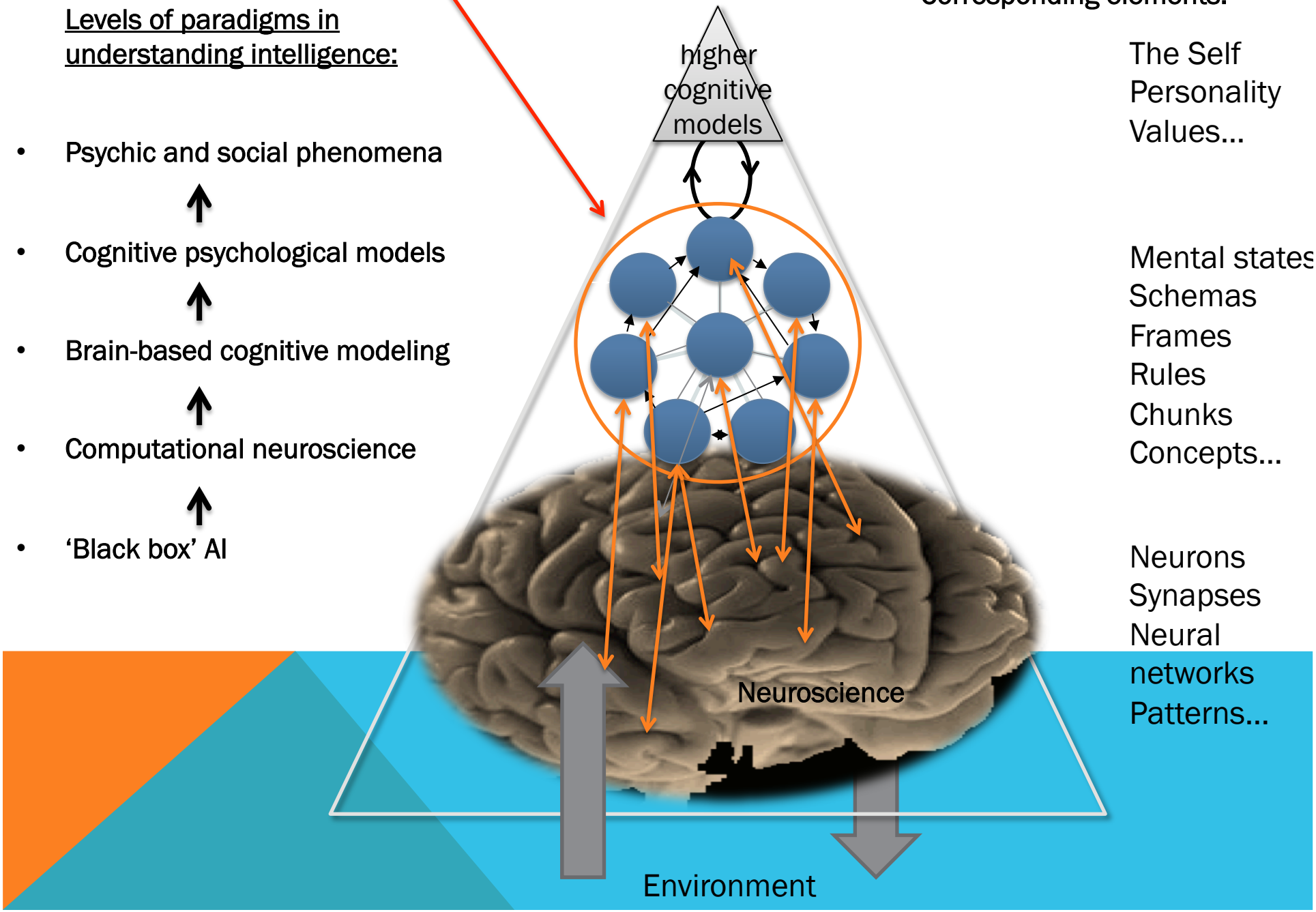
- *Lectures given via the Internet and on site in MEPhI.* You prepare and give lectures on the topics of your choice. Your trips to Russia or your time spent on the Internet will be compensated by NRNU MEPhI in the form of an honorarium.

- *Membership in the International Scientific Council of NRNU MEPhI.* You endorse the creation of the Council and occasionally provide relatively minor consulting services, steering the development of the field. You can be paid a small salary for this activity.

BICA

Levels of paradigms in understanding intelligence:

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



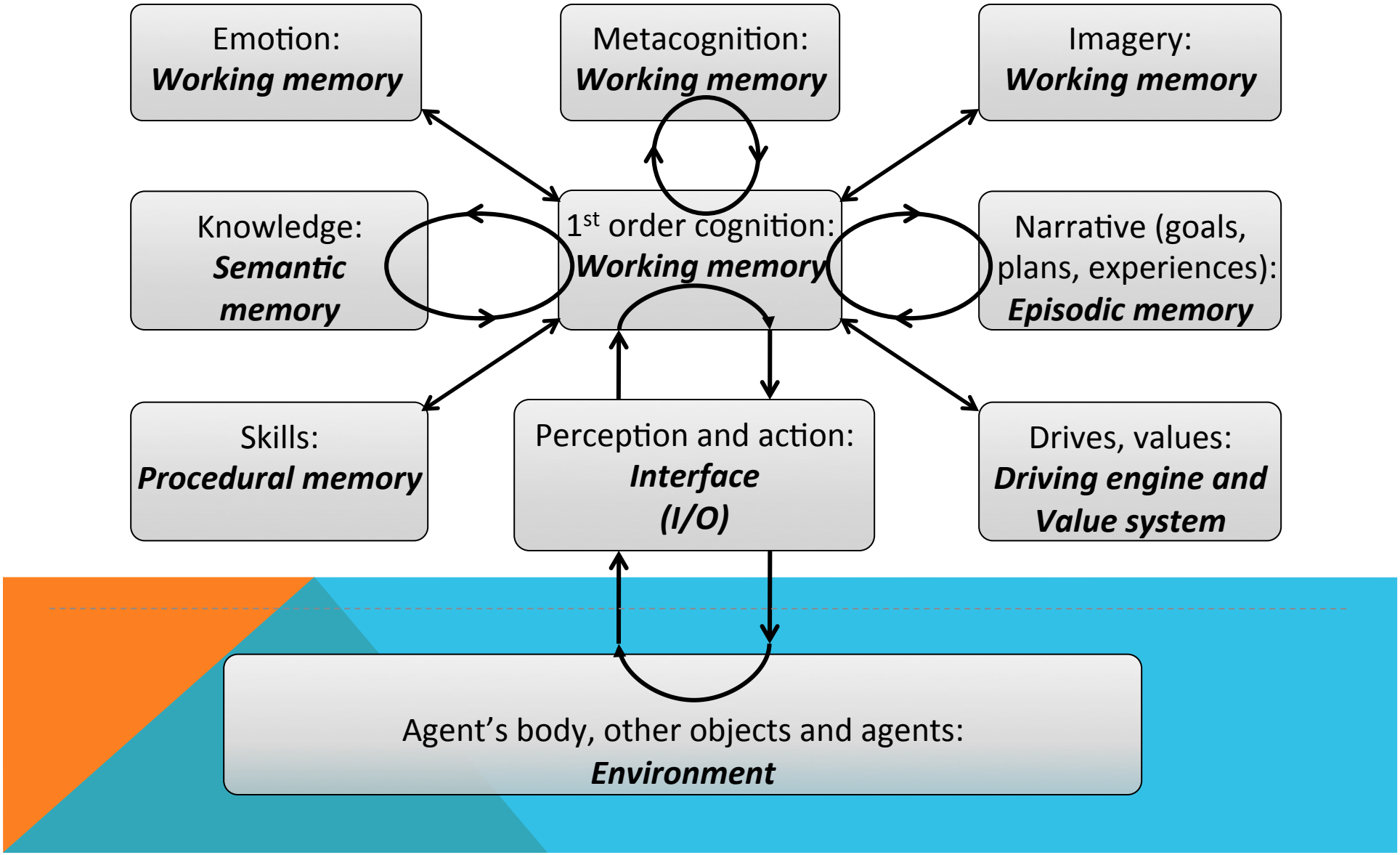
Corresponding elements:

The Self
Personality
Values...

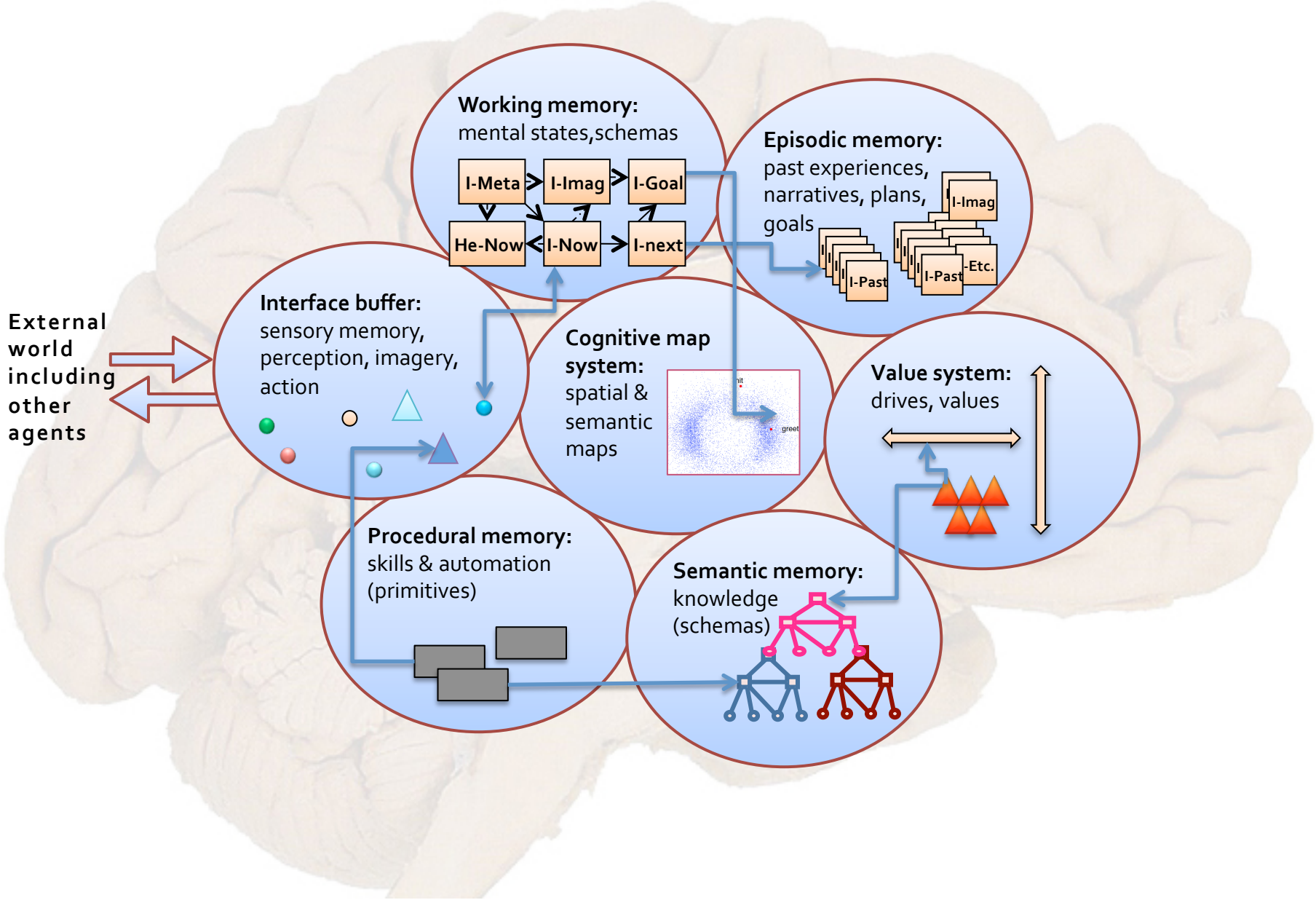
Mental states
Schemas
Frames
Rules
Chunks
Concepts...

Neurons
Synapses
Neural networks
Patterns...

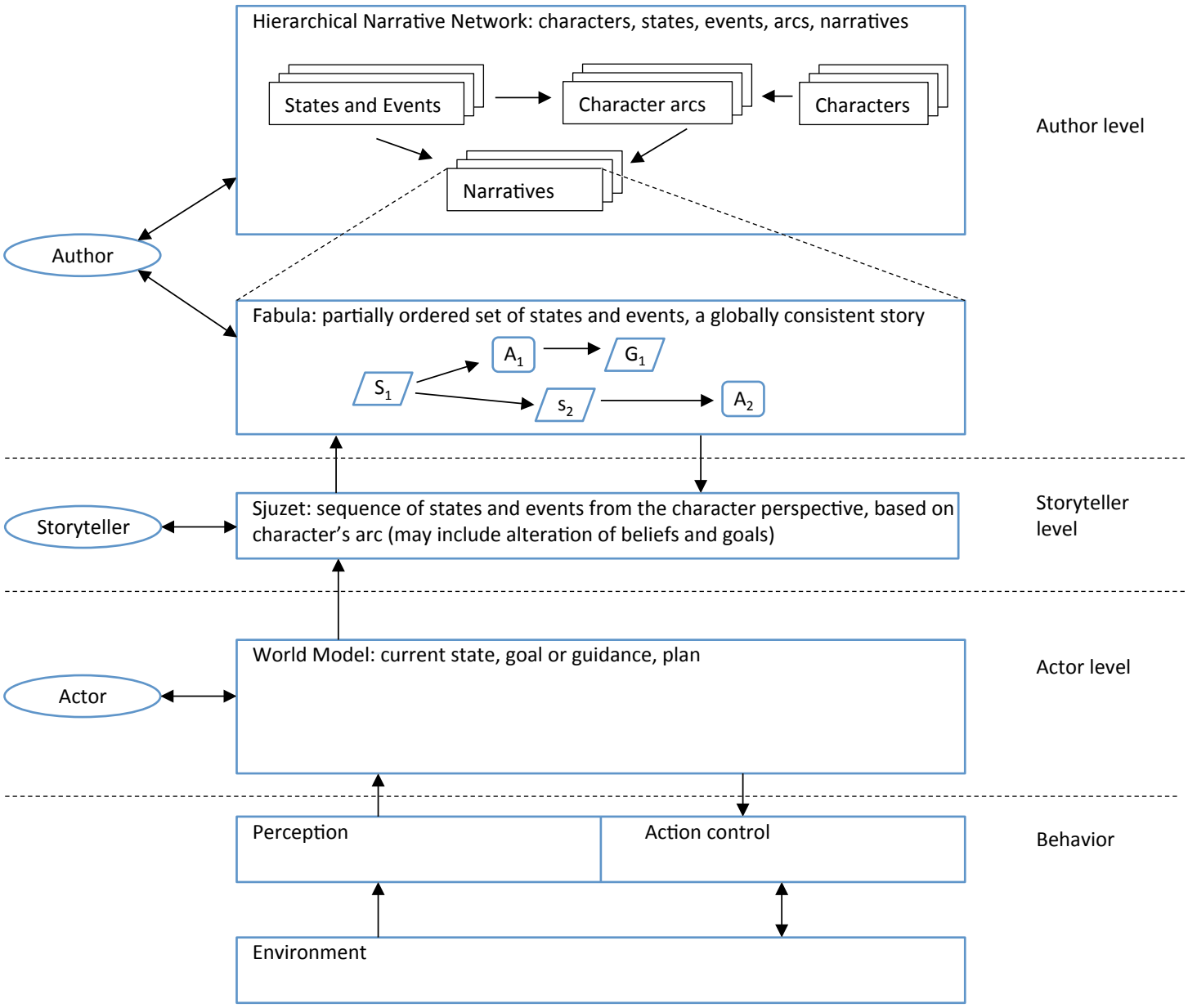
GENERIC EXTENDED COGNITIVE ARCHITECTURE



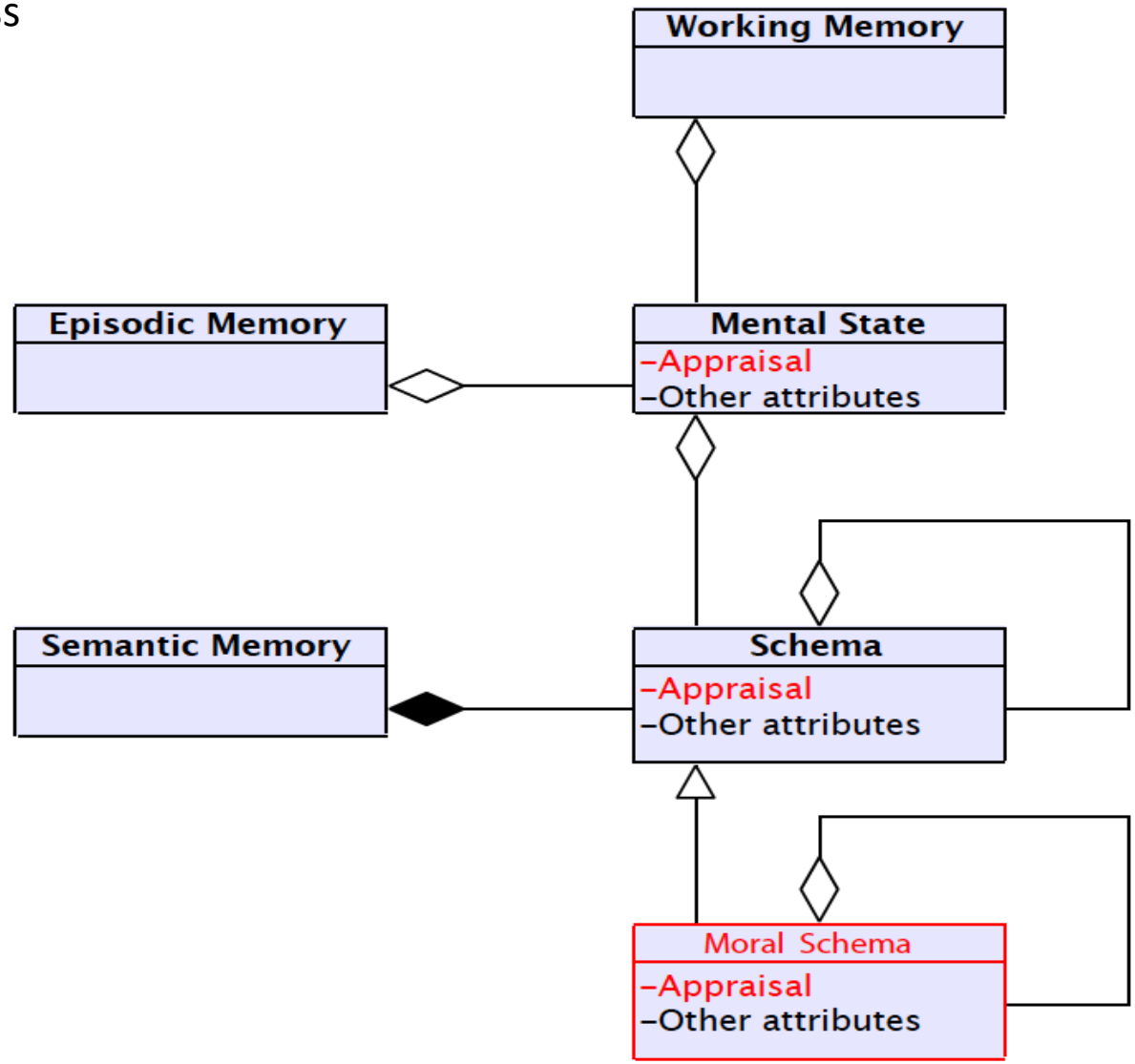
A bird's-eye view of the eBICA architecture

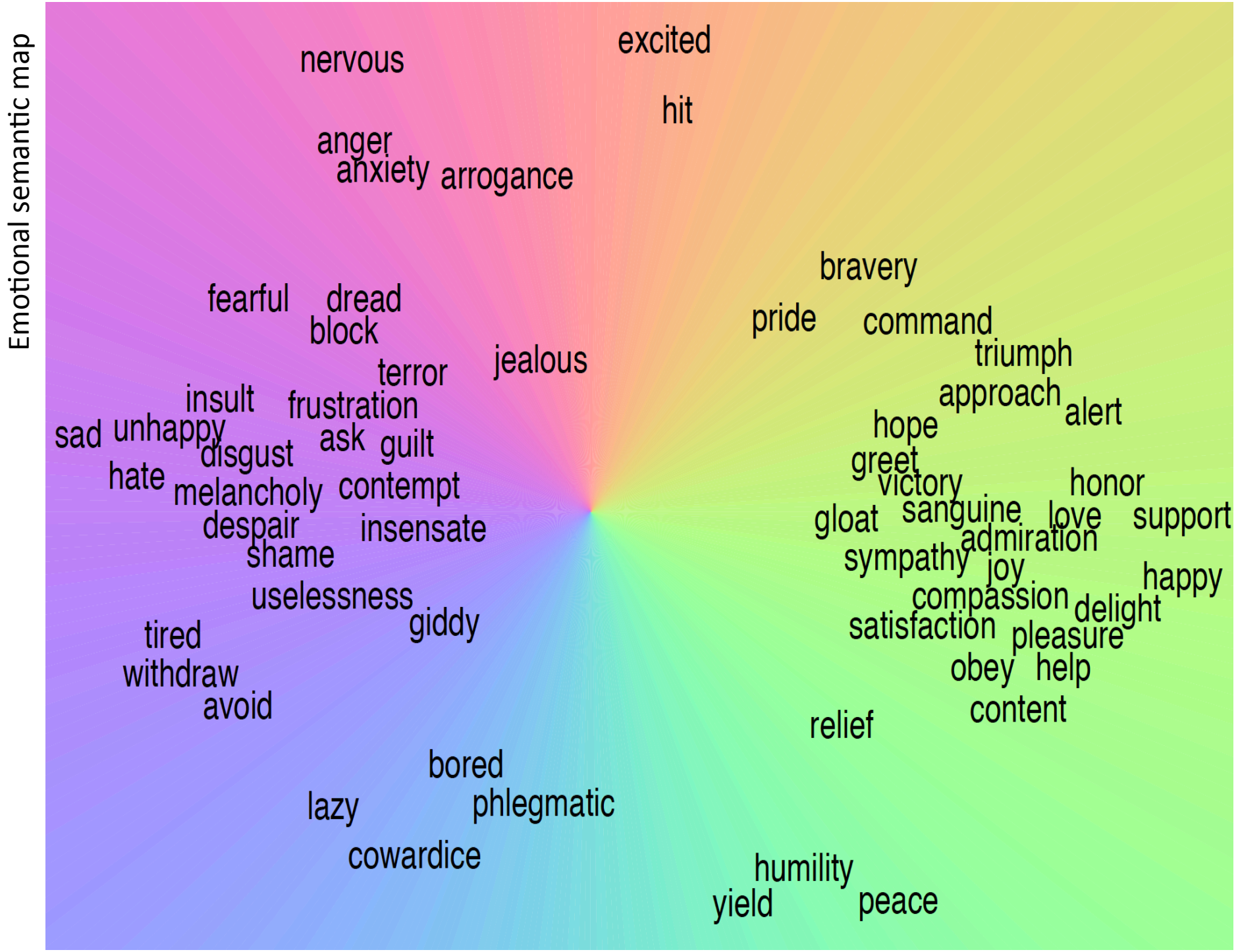


General architecture of a Character Reasoner

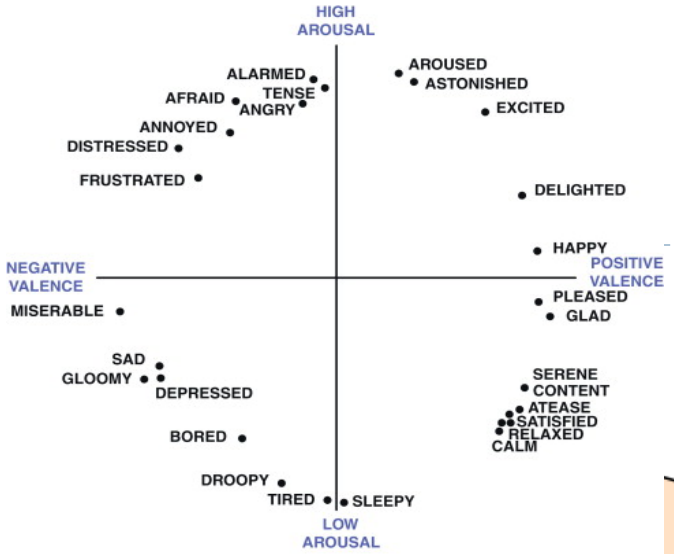


Core UML class diagram of eBICA

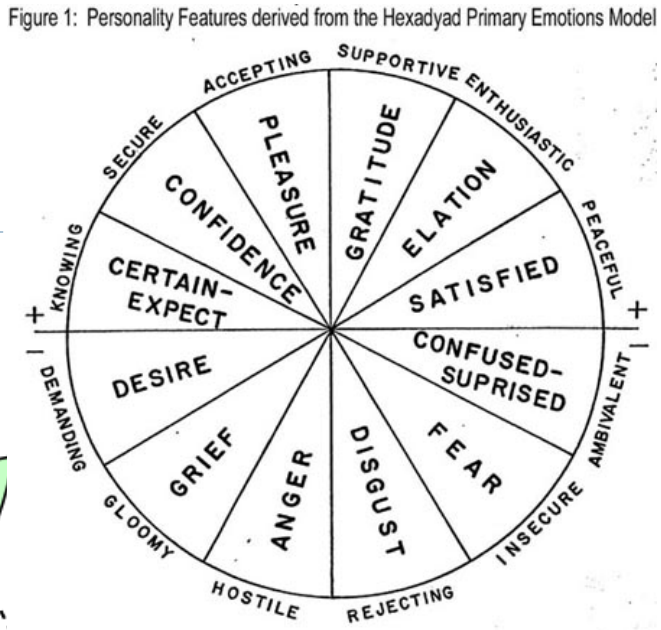


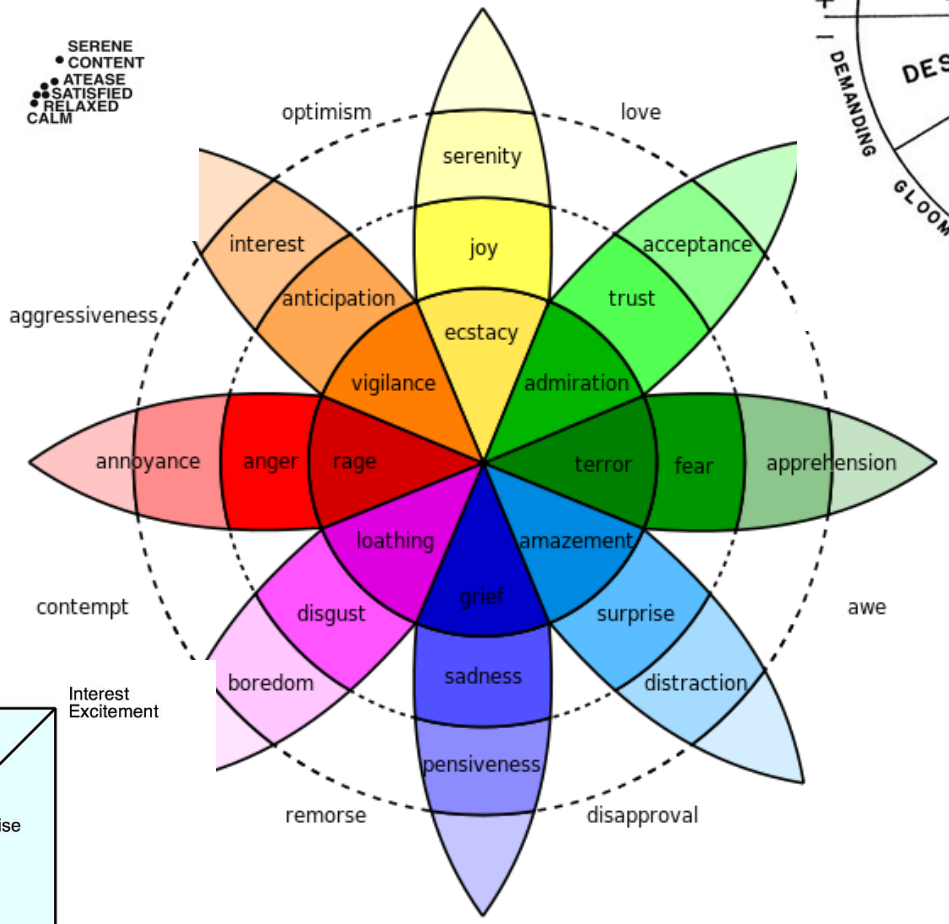

Emotional semantic map
nervous
excited
hit
anger
anxiety
arrogance
bravery
pride
command
triumph
approach
alert
hope
greet
victory
honor
gloat
sanguine
love
support
admiration
sympathy
joy
happy
compassion
delight
satisfaction
pleasure
obey
help
content
relief
fearful
dread
block
jealous
terror
insult
frustration
sad
unhappy
ask
guilt
disgust
hate
melancholy
contempt
despair
insensate
shame
uselessness
giddy
tired
withdraw
avoid
bored
lazy
phlegmatic
cowardice
humility
yield
peace

Emotion spaces



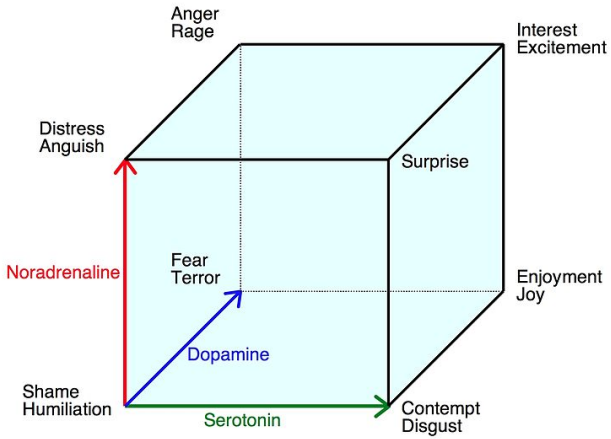
(Russell 1980)





← Plutchik

Cube of emotions (Lövheim 2012)



Other well-known models include:

- PAD (ANEW)
- EPA
- Semantic differential

The core equations of emotionally biased cognition and decision making

The virtual actor used in this study was built based on the eBICA architecture (Samsonovich, 2013). It's cognitive processes are described in terms of appraisals A and likelihoods L:

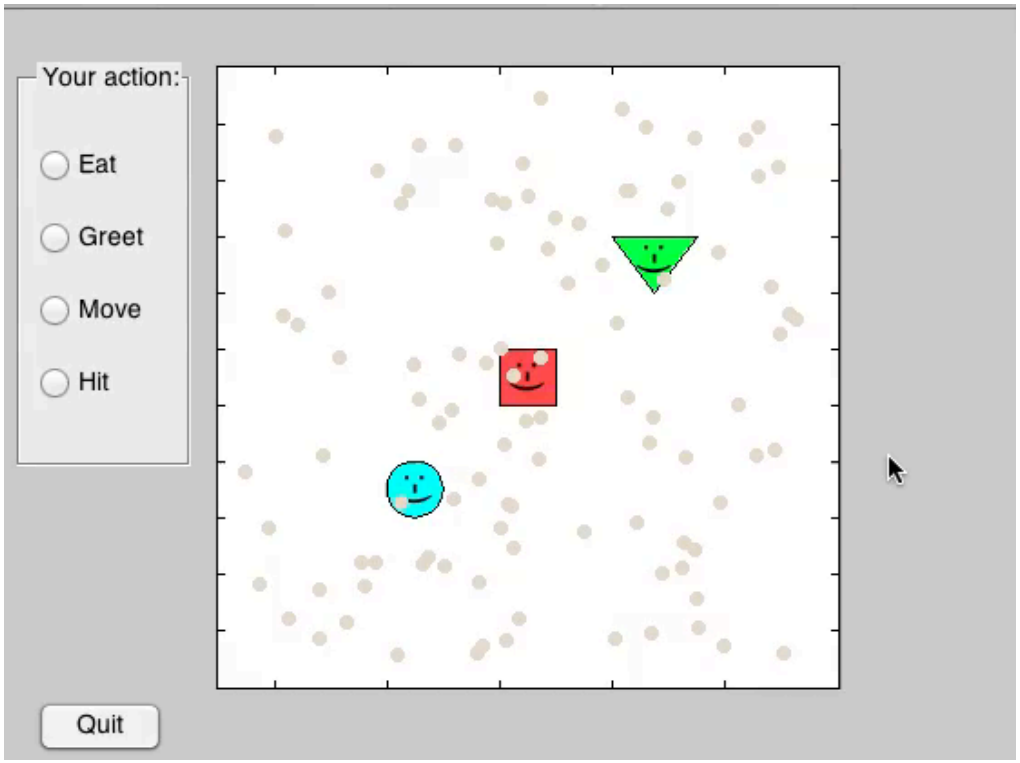
$$A_{recipient}^{t+1} = (1-r)\, A_{recipient}^{t} + r\, A_{action}$$

$$A_{agent}^{t+1} = (1-r)\, A_{agent}^{t} + r\, A^{*}_{action} \quad (1)$$

$$L_{action} \sim [\mathrm{Re}(A_{action}(A^{*}_{agent} + A_{recipient}))]_{+}. \quad (2)$$

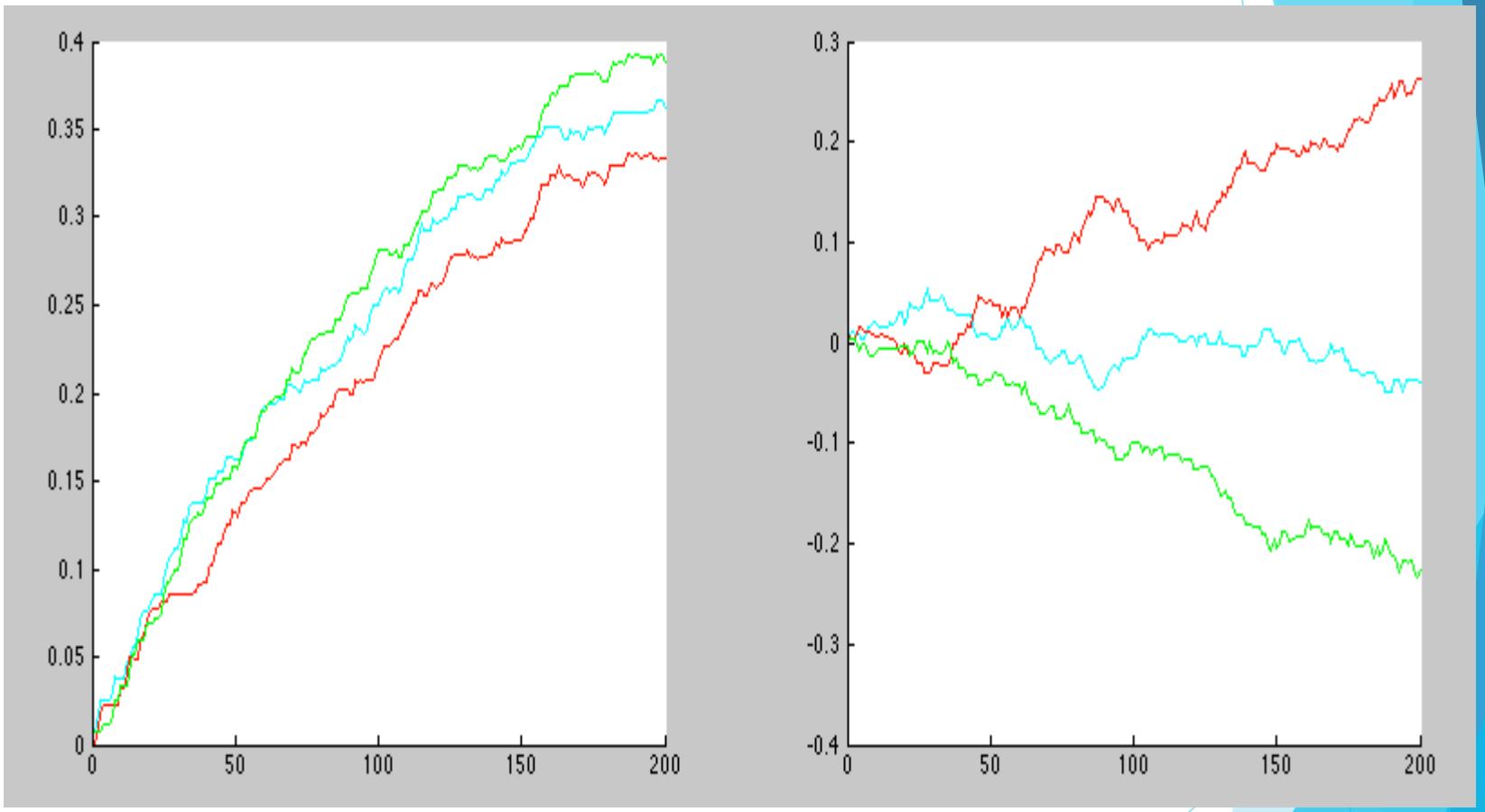
Here the appraisal A of an object or an action is given by a complex number with its real part representing the valence and its imaginary part representing a mixture of dominance and arousal (again, this is only a first approximation of a theory: more detailed models should include more than two components). The time t is discrete, and the constant r is a model parameter.

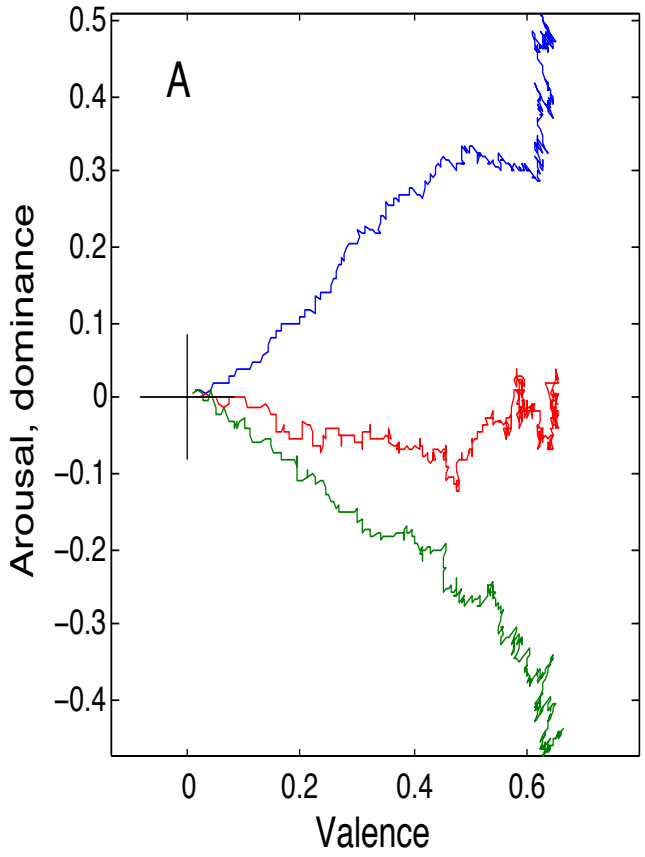
A simple videogame engaging a small group of actors in social interactions

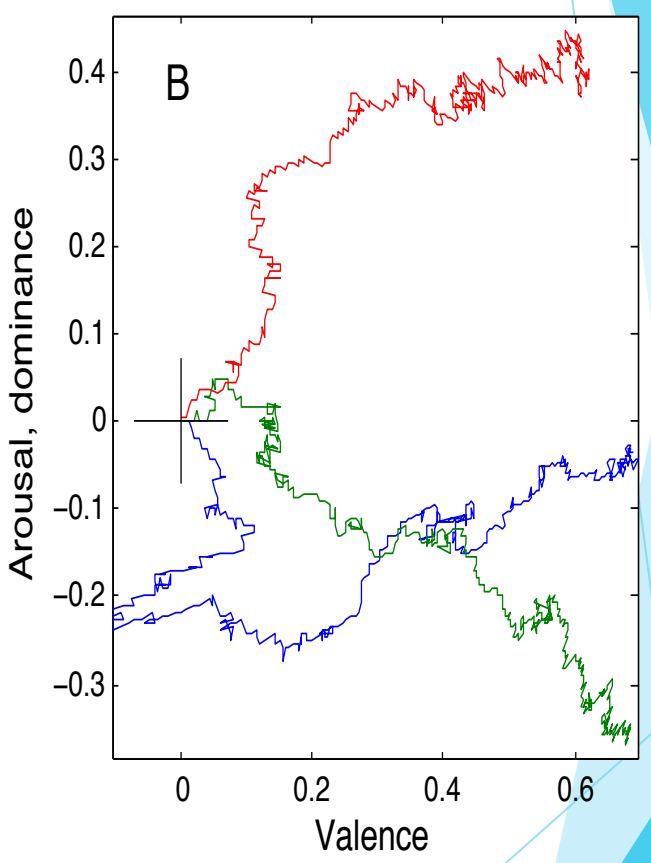


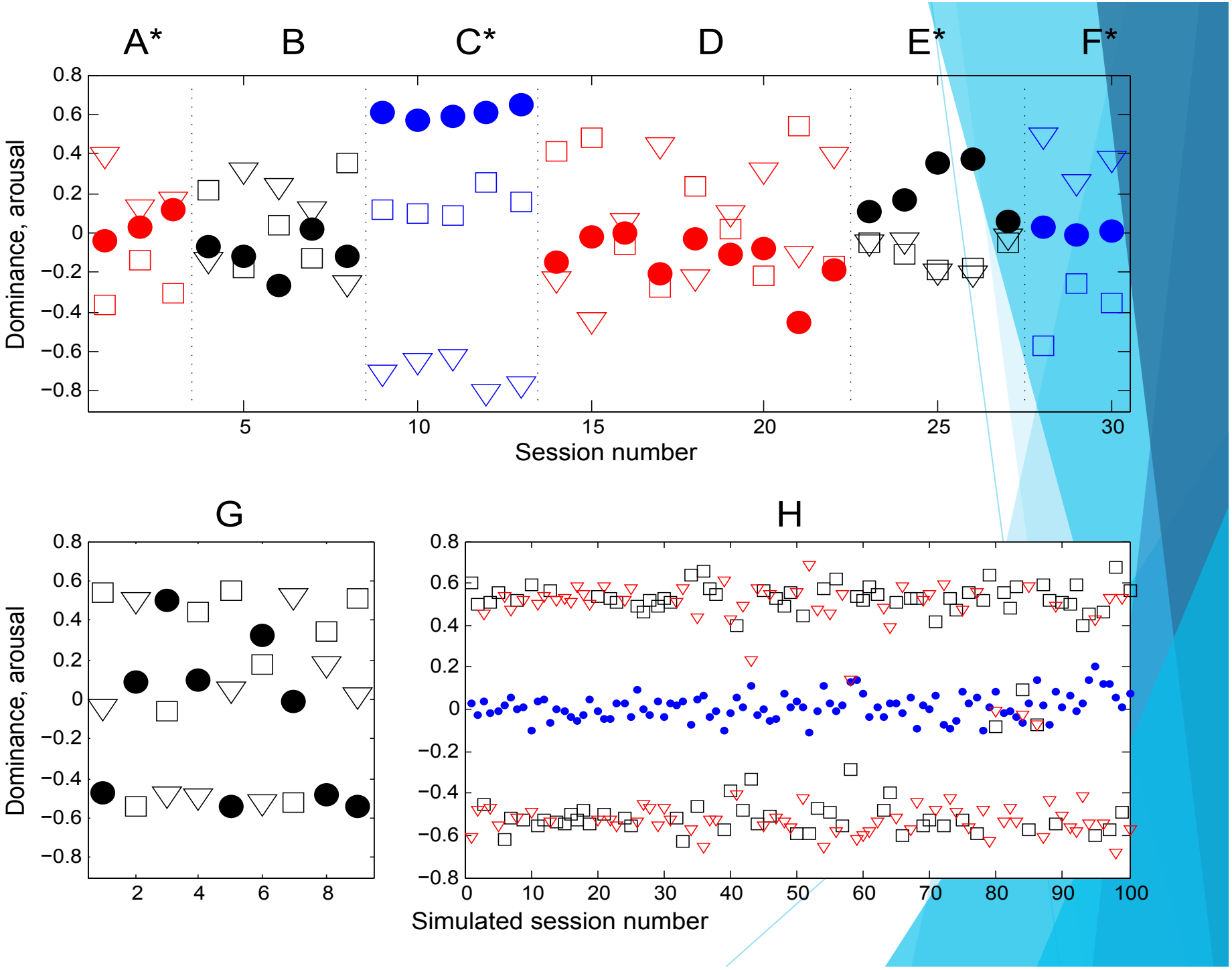
valence

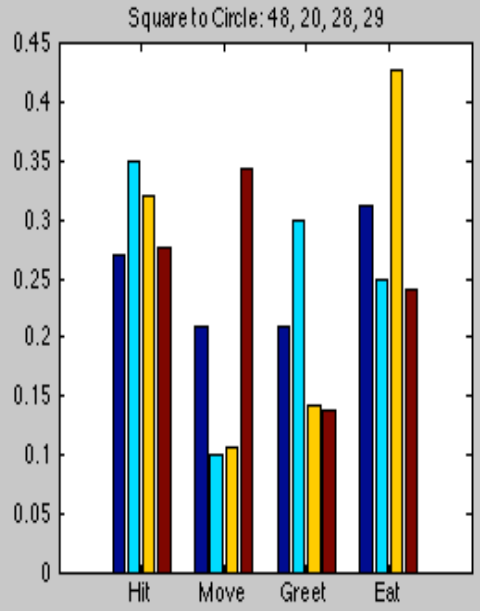
dominance

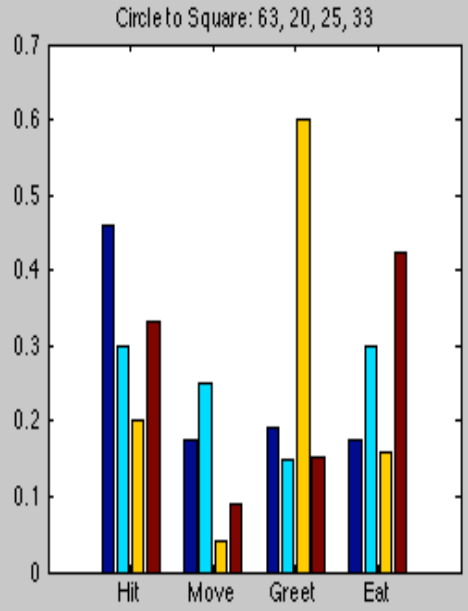


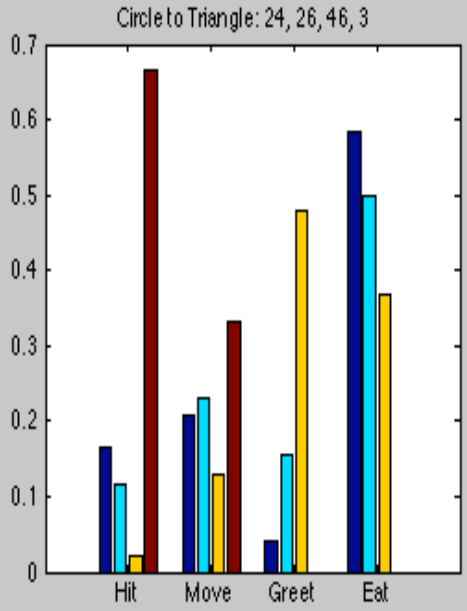

A
Arousal, dominance
Valence
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
0
0.2
0.4
0.6

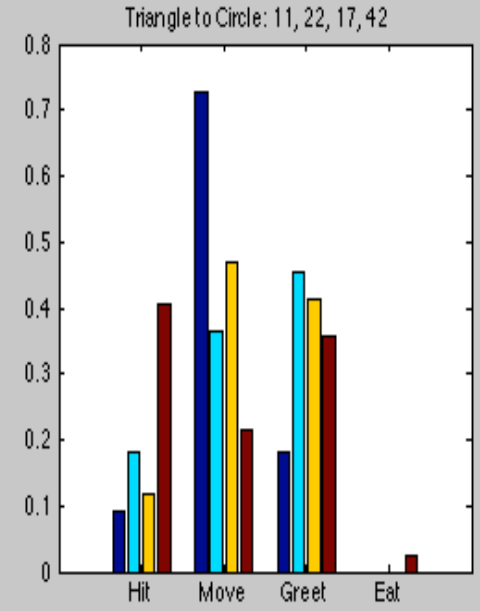

B
Arousal, dominance
Valence
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
0
0.2
0.4
0.6

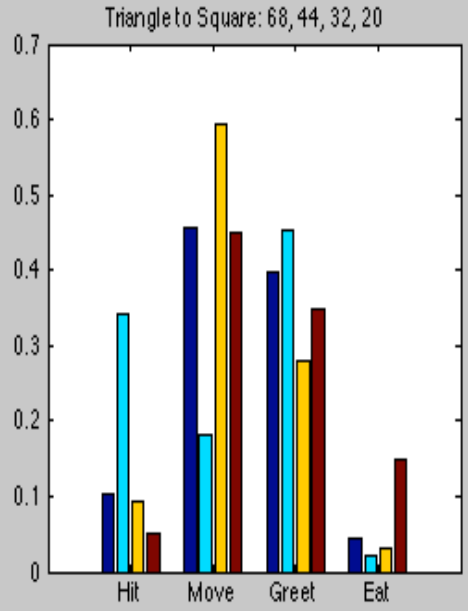

A*
B
C*
D
E*
F*
Dominance, arousal
Session number
G
H
Simulated session number

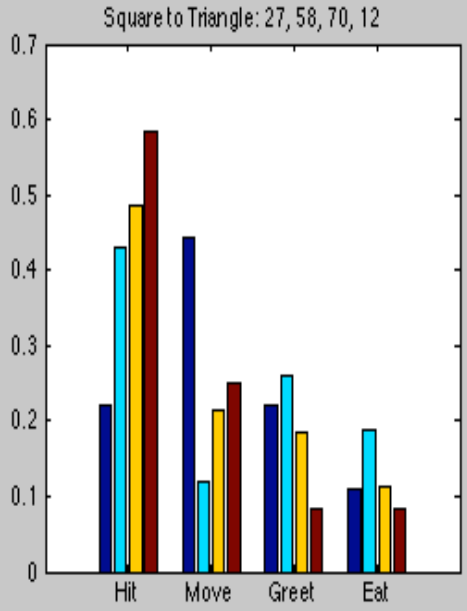

Square to Circle: 48, 20, 28, 29
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Hit
Move
Greet
Eat


Circle to Square: 63, 20, 25, 33
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Hit
Move
Greet
Eat

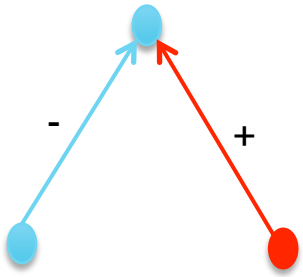

Circle to Triangle: 24, 26, 46, 3
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Hit
Move
Greet
Eat

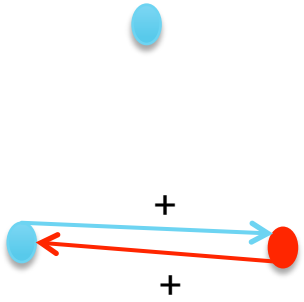

Triangle to Circle: 11, 22, 17, 42
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Hit
Move
Greet
Eat


Triangle to Square: 68, 44, 32, 20
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Hit
Move
Greet
Eat


Square to Triangle: 27, 58, 70, 12
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Hit
Move
Greet
Eat

Examples of noticeable patterns suggest the need to define "moral schemas"

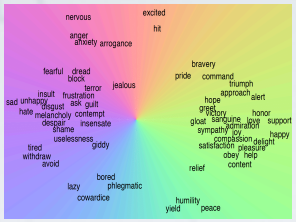




They become building blocks for narrative networks

Adding moral schemas:

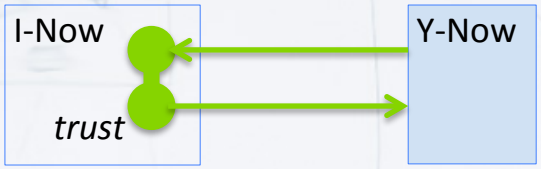
- An instantiated **moral schema of a social relationship** becomes a “character on its own” and a “pattern recognizer” that stabilizes mutual appraisals
- It defines **feelings** about actors that replace appraisals
- As a character, it has motives that generate goals
- When the pattern is no longer recognized, the moral schema tries to address the problem
- This mechanism could naturally explain the clustering of social emotions



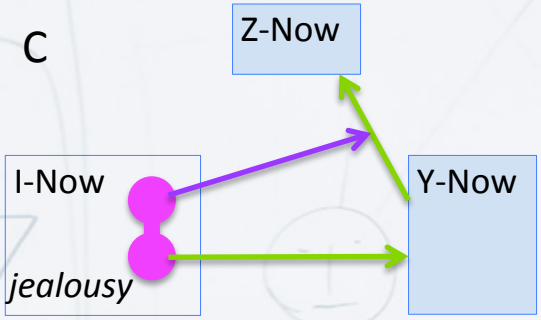
Possible models of social emotions

A

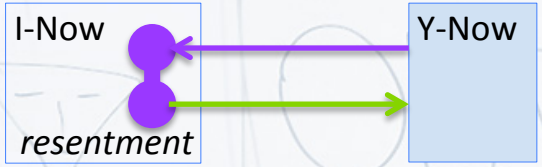
B



C



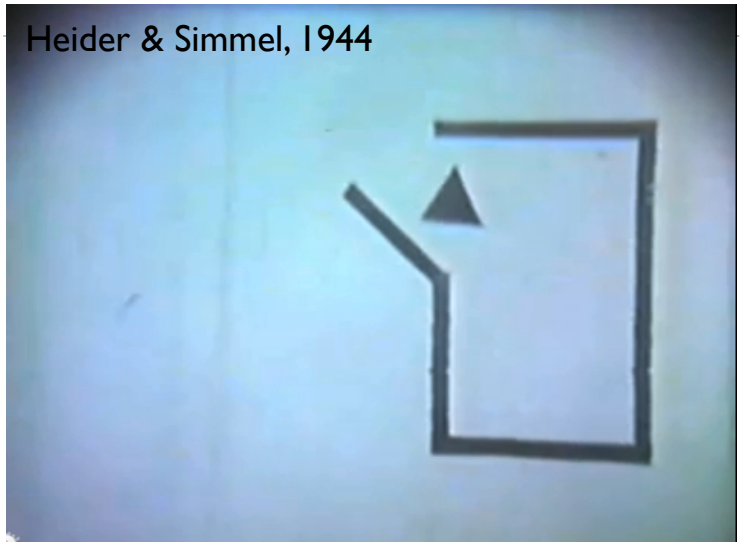
D



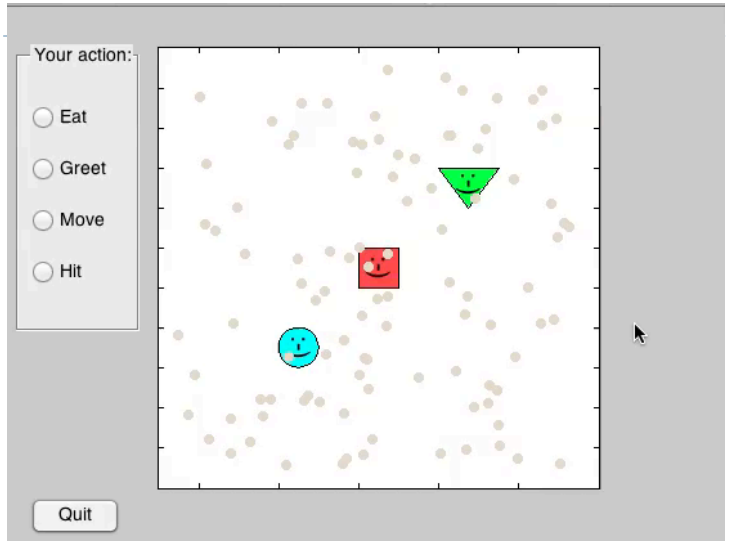
How can we validate these ideas by observing and simulating social interactions in a small group of actors?

Experiments involving humans and machines

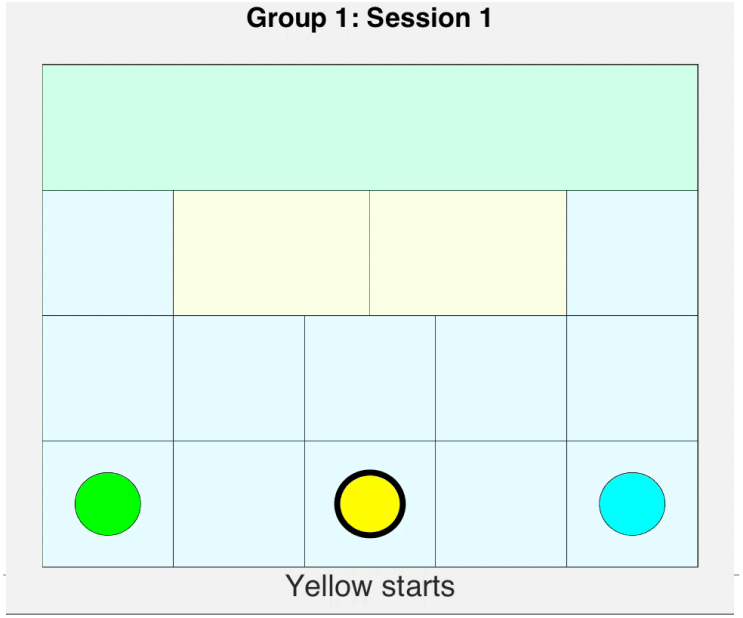
A



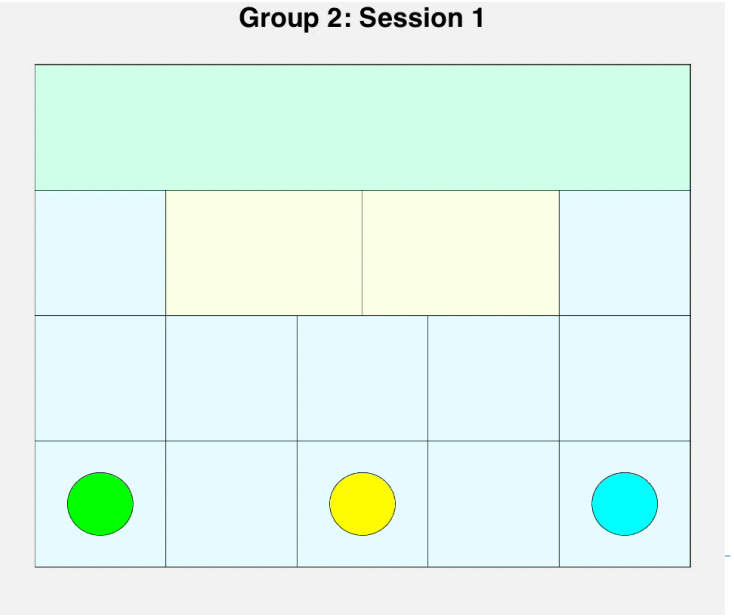
B



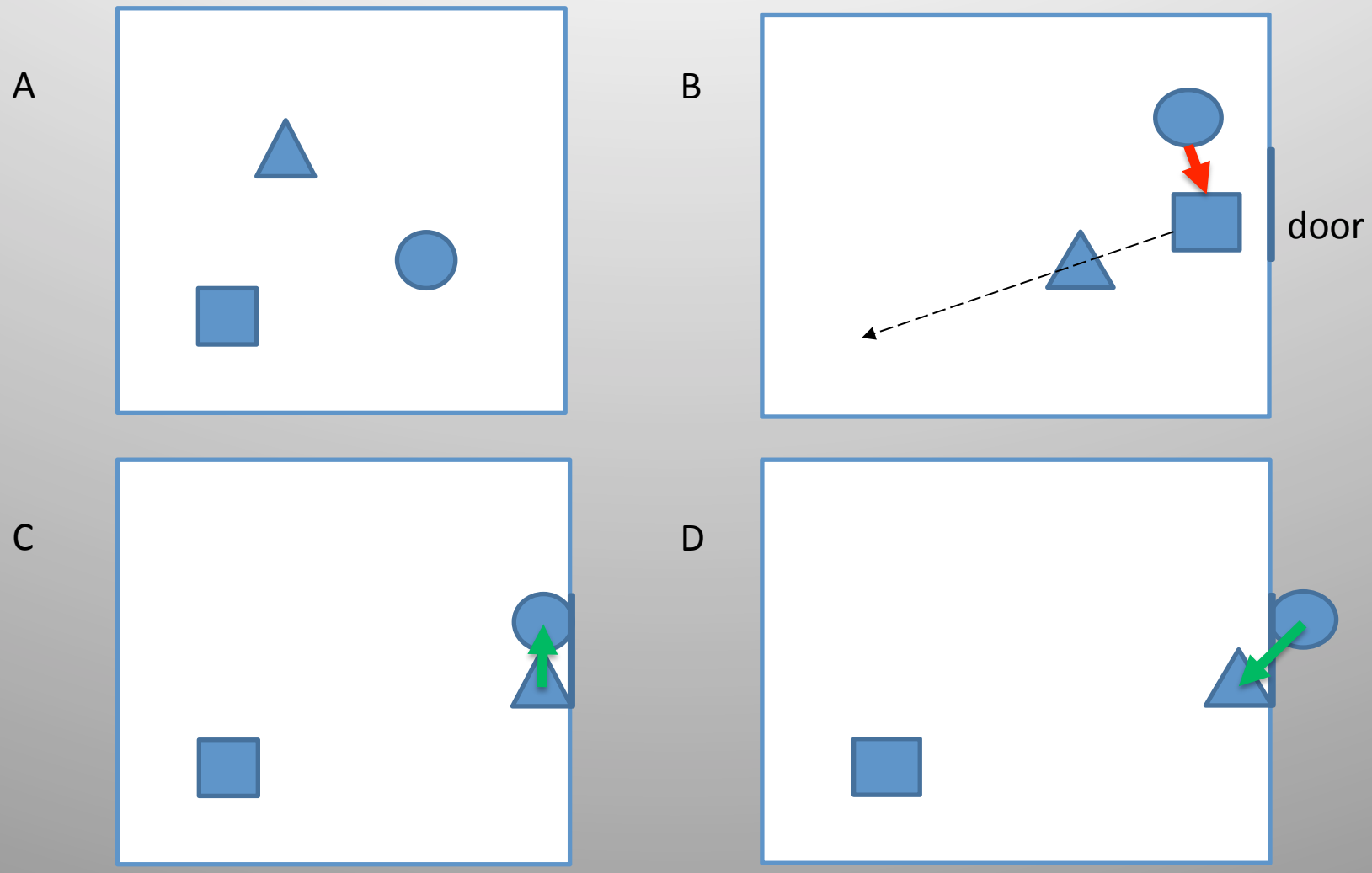
C



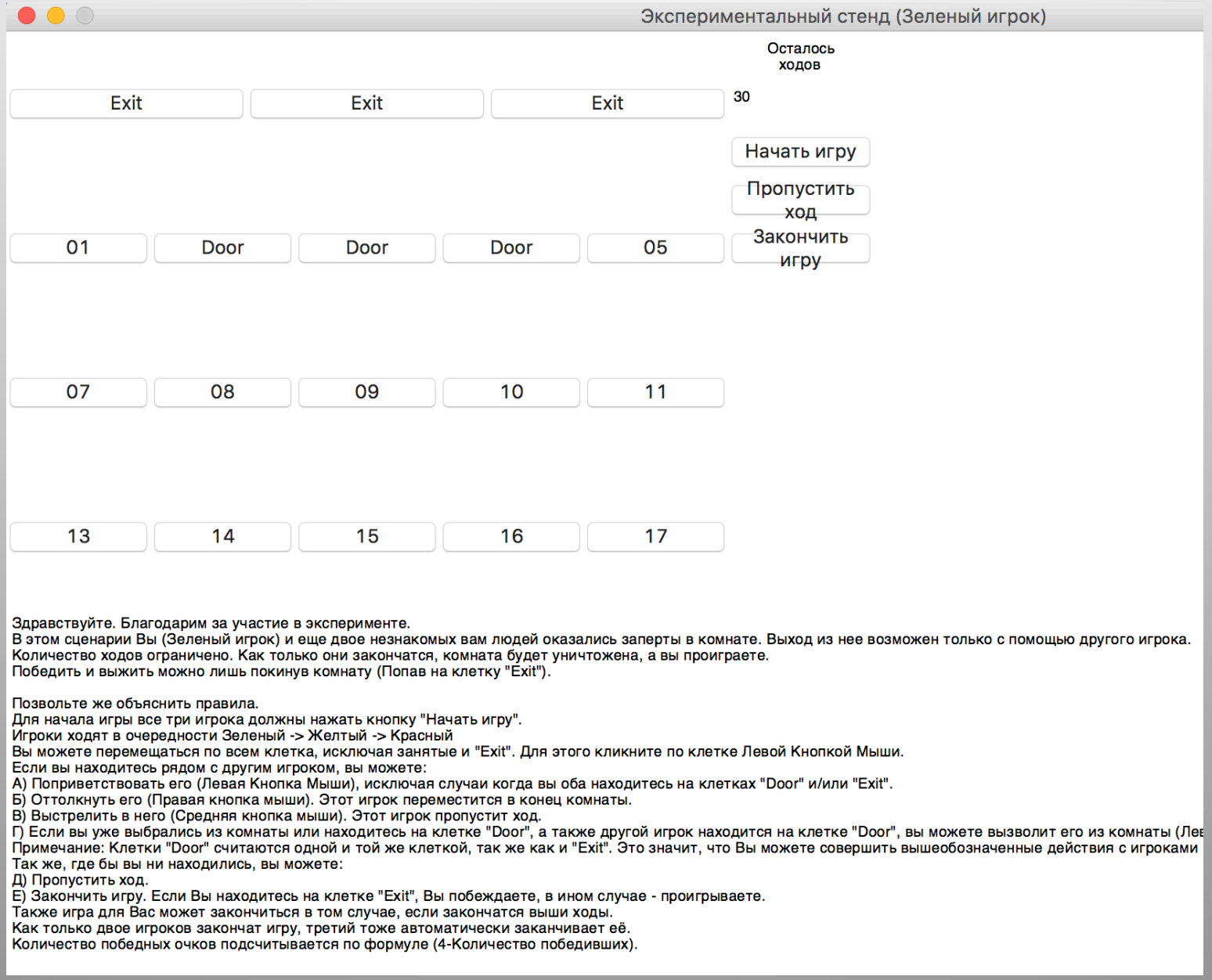
D



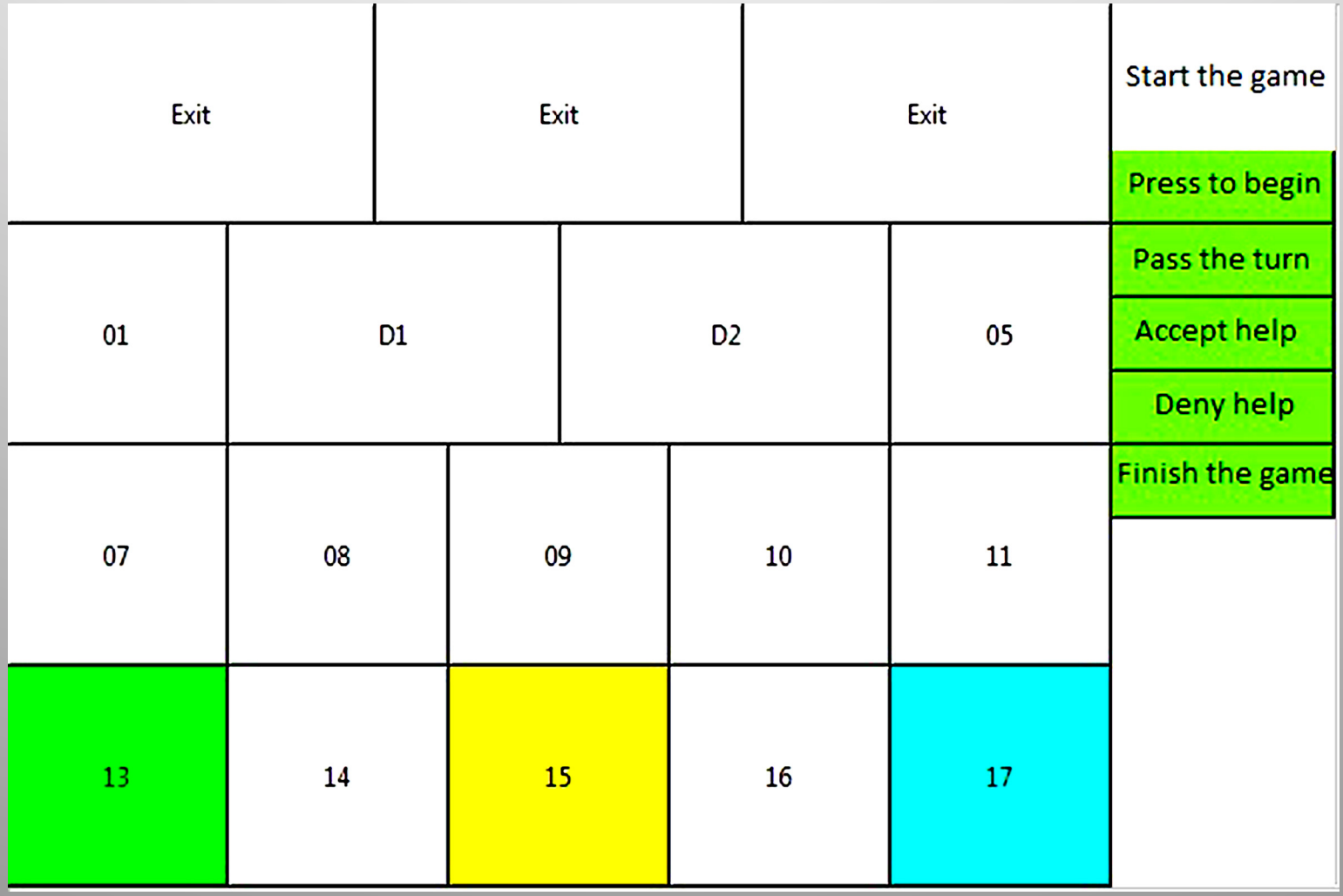
“Russian elevator story”: A challenge for machines



Work in progress..



A user interface snapshot



Predictions-hypotheses:

- One couple takes over
- Wrong behavior changes relations

Actual outcome (preliminary):

- A hierarchy develops: one always escapes first
- One hierarchy switches to another

Preliminary results

Order of escapes

- CGY
- CGY
- CYG
- YCG
- CGY
- GYC
- Y
- YGC
- YGC
- GCY
- CGY
- YGC
- GC
- YCG
- CYG
- CGY
- CY
- CYG

Frequencies of actions

--------shots

0	2	3	5
1	0	4	5
3	3	0	6
4	5	7	16
3	6	7	

--------greets

0	10	8	18
6	0	9	15
1	4	0	5
7	14	17	38
16	9	13	

--------kicks

0	10	8	18
4	0	2	6
3	7	0	10
7	17	10	34
14	11	9	

--------helps and greets

0	12	11	23
10	0	16	26
7	12	0	19
17	24	27	68
22	18	28	

--------kicks and shots

0	12	11	23
5	0	6	11
6	10	0	16
11	22	17	50
17	17	16	

--------helps minus kicks

0	−8	−5	−13
0	0	5	5
3	1	0	4
3	−7	0	−4
−8	−2	6	

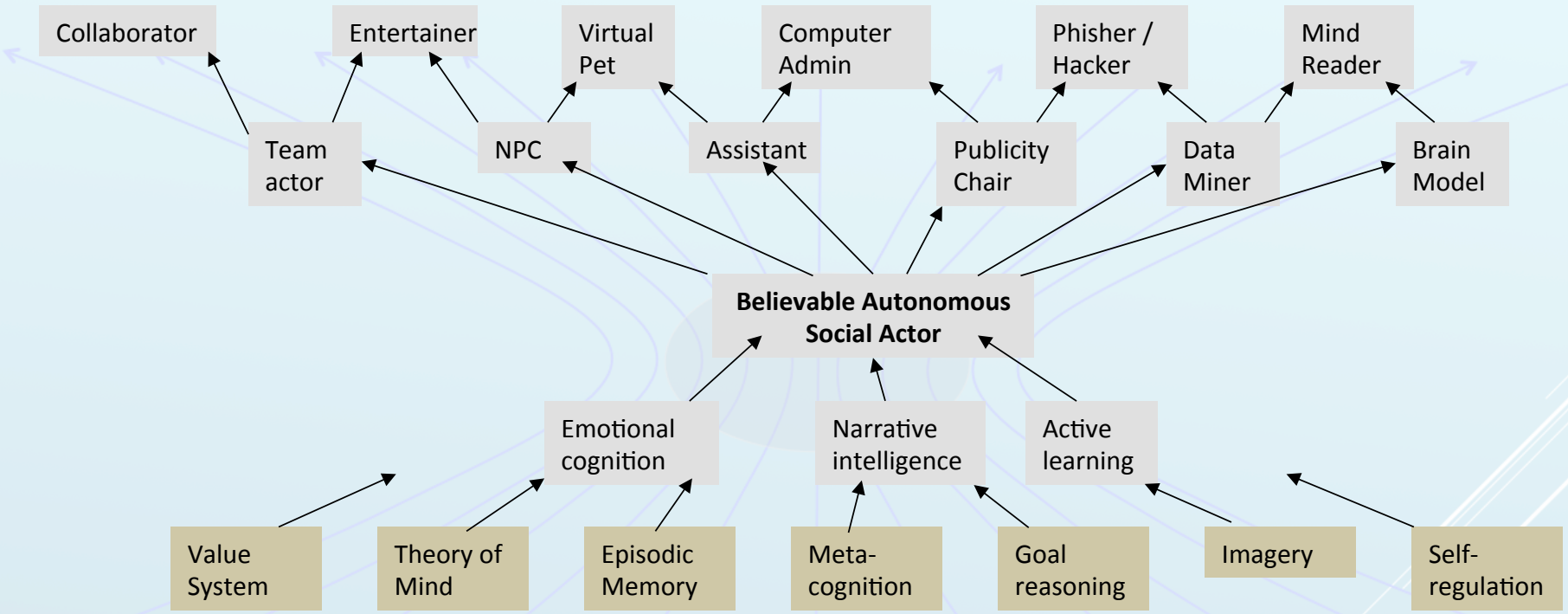
--------helps

0	2	3	5
4	0	7	11
6	8	0	14
10	10	10	30
6	9	15	

So what?

- Emotional intelligence may be relatively simple to capture with a mathematical model
- This core functional unit can be “mounted” on top of virtually any given intelligent agent, potentially making it “believable”
- But the question is how to define the goal: the task or challenge that if solved, takes us to the next level? This is a challenge..
- The described Turing-test-like videogame scenario may be an answer

A look into the future:
Possible structure of a breakthrough in AI



What do we need right now for this to happen?

- Believable character reasoning
- Social-emotional intelligence
- Human-like learning capabilities
- Useful, precise metrics and tests (challenges)

BICA Society

Biologically Inspired Cognitive Architectures Society

BICA Events

BICA Conference Series

- BICA 2018: TBA
- BICA 2017: Moscow, Russia (sponsored by the Russian Science Foundation). Tentative dates: TBA. **Chair:** Alexei Samsonovich. Participants: Mike Sellers,..
- BICA 2016: New York City, NY, USA. Dates: Saturday, July 16, to Tuesday, July 19, 2016. BICA 2016 will be hosted as a part of the unified HLAI Framework event, also including AGI-2016 (www.agi-conf.org), NeSy-2016 (www.neural-symbolic.org), possibly AIC-2016 and more, co-located and immediately following IJCAI-2016 (http://ijcai-16.org). **General Chair of HLAI 2016:** Tarek R. Besold. BICA 2016 Chair: Alexei Samsonovich
- BICA 2015 (November 6-8): Lyon, France. **Chair:** Amélie Cordier (amelie.cordier@liris.cnrs.fr). General Program Chairs: Alexei Samsonovich and Olivier Georgeon.
- BICA 2014 (November 7-9): MIT, Boston, MA. **Chair:** Paul Robertson (paulr@dollabs.com, drpaulrobertson@gmail.com). Co-Chairs: Patrick H. Winston, Howard Shrobe, and Alexei Samsonovich