

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)
РОССИЙСКАЯ АКАДЕМИЯ НАУК
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ СИСТЕМНЫХ ИССЛЕДОВАНИЙ РАН
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»
МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
ТРОИЦКИЙ ИНСТИТУТ ИННОВАЦИОННЫХ И ТЕРМОЯДЕРНЫХ ИССЛЕДОВАНИЙ (ТРИНИТИ)
РОССИЙСКАЯ АССОЦИАЦИЯ НЕЙРОИНФОРМАТИКИ
РОССИЙСКАЯ АССОЦИАЦИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

XXI МЕЖДУНАРОДНАЯ НАУЧНО-ТЕХНИЧЕСКАЯ КОНФЕРЕНЦИЯ

НЕЙРОИНФОРМАТИКА-2019

СБОРНИК НАУЧНЫХ ТРУДОВ

ЧАСТЬ 2

- **ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ**
- **КОГНИТИВНЫЕ НАУКИ И ИНТЕРФЕЙС
“МОЗГ-КОМПЬЮТЕР”**
- **АДАПТИВНОЕ ПОВЕДЕНИЕ И ЭВОЛЮЦИОННОЕ
МОДЕЛИРОВАНИЕ**
- **НЕЙРОМОРФНЫЕ ВЫЧИСЛЕНИЯ, ГЛУБОКОЕ
ОБУЧЕНИЕ**
- **ПРИКЛАДНЫЕ НЕЙРОСЕТЕВЫЕ СИСТЕМЫ**
- **ТЕОРИЯ НЕЙРОННЫХ СЕТЕЙ, НЕЙРОСЕТЕВЫЕ
ПАРАДИГМЫ И АРХИТЕКТУРЫ**
- **НЕЙРОБИОЛОГИЯ И НЕЙРОБИОНИКА**

МОСКВА
2019

УДК 001(06)+004.032.26(06)
ББК 72я5+32.818я5
М 82

**XXI МЕЖДУНАРОДНАЯ НАУЧНО-ТЕХНИЧЕСКАЯ КОНФЕРЕНЦИЯ
"НЕЙРОИНФОРМАТИКА-2019"** : сборник научных трудов. В 2-х частях.
Часть 2. – Москва : МФТИ, 2019. 213 с. : ил.
ISBN 978-5-89155-323-1 (Ч. 2)
ISBN 978-5-89155-321-7

Сборник научных трудов содержит доклады, включенные в программу XXI Международной научно-технической конференции «НЕЙРОИНФОРМАТИКА-2019», проходившей в г. Москве 7–11 октября 2019 г. Тематика конференции охватывает широкий круг вопросов: искусственный интеллект, глубокое обучение, когнитивные науки и интерфейс «мозг-компьютер», методические вопросы нейроинформатики, теории нейронных сетей, нейробиологии, модели адаптивного поведения и когнитивные исследования, нейросетевые парадигмы и приложения нейроинформатики.

УДК 001(06)+004.032.26(06)
ББК 72я5+32.818я5

ISBN 978-5-89155-323-1 (Ч. 2)
ISBN 978-5-89155-321-7

© Федеральное государственное автономное
образовательное учреждение высшего
образования «Московский физико-
технический институт (национальный
исследовательский университет)», 2019

Е. Е. ВИТЯЕВ

Институт математики им. С.Л. Соболева, Новосибирск
vityaev@math.nsc.ru

СОЗНАНИЕ КАК КОМПЛЕКСНОЕ ОТРАЖЕНИЕ ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ ВНЕШНЕГО МИРА*

В предыдущих работах была решена проблема статистической двусмысленности и определены максимально специфические причинно-следственные связи, вывод по которым непротиворечив. В данной работе аргументируется следующая гипотеза: мозг делает все возможные выводы по воспринимаемым причинно-следственным связям, создавая непротиворечивую модель воспринимаемого мира, проявляющуюся как сознание. Показано, что предложенная модель сознания является вариантом теории сознания G. Tononi, основанной на интегрированной информации.

Ключевые слова: категоризация, естественная классификация, естественные понятия, интегрированная информация.

Введение

G. Tononi [1–3] определяет сознание как первичное понятие, которое обладает следующими феноменологическими свойствами: composition, information, integration, exclusion. Для более точного определения этих свойств G. Tononi [1] вводит понятие интегрированной информации: «интегрированная информация, характеризующая редукцию неопределенности, – это информация, генерируемая системой, приходящей в некоторое состояние через причинное взаимодействие между ее частями, которая превосходит информацию, генерируемую независимо ее частями самими по себе». В терминах интегрированной информации феноменологические свойства формулируются следующим образом.

- 1) composition – elementary mechanisms (causal interactions) can be combined into higher order ones;
- 2) information – only mechanisms that specify ‘differences that make a difference’ within a system count;
- 3) integration – only information irreducible to non-interdependent components counts;
- 4) exclusion – only maxima of integrated information count.

* Работа выполнена при частичной поддержке РФФИ, грант № 19-01-00331 а.

Г. Топони рассматривает эти свойства как внутренние свойства системы. Интегрированная информация у Г. Топони рассматривается как система циклических причинных связей. Однако он не определяет, что же отражает интегрированная информация помимо феноменологических свойств.

В данной работе нами выдвигается гипотеза о том, что мозг с помощью интегрированной информации настраивается на восприятие «естественных» объектов внешнего мира. Мы рассматриваем эти свойства не как внутренние свойства системы, а как способность системы отражать комплексы причинных связей объектов внешнего мира, а сознание – как способность комплексного иерархического отражения «естественной» классификации внешнего мира.

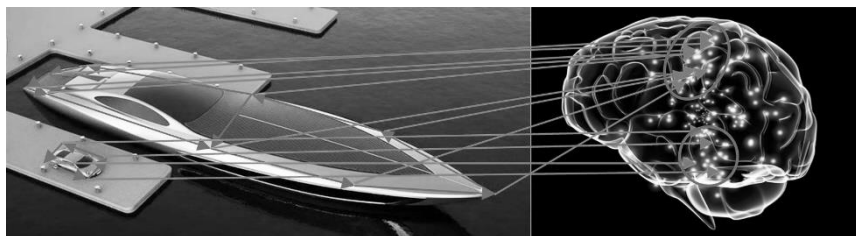


Рис. 1. Отражение мозгом причинных связей между свойствами объектов

Рассмотрим процесс отражения причинных связей (рис. 1). Он включает:

- 1) объекты внешнего мира (машина, лодка, причал), относящиеся к некоторым «естественным» классам;
- 2) процесс отражения мозгом свойств объектов и связывающих их причинных связей;
- 3) объединение возбужденных нейронов мозга в системы, обозначенные овалами.

В теории Г. Топони рассмотрен только третий пункт процесса отражения. Мы покажем, что интегрированная информация может рассматриваться не только как формализация феноменологических свойств, но и как комплексное иерархическое отражение «естественной» классификации внешнего мира, при определенной ее модификации.

Причинные связи и «естественная» классификация

Причинность является следствием физического детерминизма: «для любой изолированной физической системы некоторое ее состояние определяет все последующие состояния» [4]. Рассмотрим автомобильную аварию [4]. В чем её причина? Это может быть состояние дорожного покрытия

тия, его влажность, положение солнца относительно водителя, безрассудное вождение, психологическое состояние водителя, неисправность тормозов и т.д. Очевидно, что в этом случае нет определенной причины.

В философии науки причинность сводится к предсказанию и объяснению. «Причинно-следственная связь означает предсказуемость... в том случае, если известна вся предыдущая ситуация, событие может быть предсказано..., если даны все факты и законы природы, связанные с этим событием» [4]. Понятно, что знать все факты, число которых в случае аварии потенциально бесконечно, и все законы никто не может. Кроме того, человек и животные узнают законы путем обучения (индуктивного вывода). Поэтому причинность сводится к предсказанию с помощью индуктивно-статистического вывода, когда предсказание логически выводятся из фактов и статистических законов с некоторой вероятностной оценкой.

Причинно-следственные связи, обнаруженные на реальных данных или в результате обучения, сталкиваются с проблемой статистической двусмысленности – из них могут быть выведены противоречивые предсказания [5–6]. Чтобы избежать этой двусмысленности, Гемпель ввел требование максимальной специфичности [5–6], состоящее в том, чтобы статистические законы включали максимум информации, связанные с прогнозируемым свойством. Следуя Гемпелю, мы определили максимально специфические правила, для которых удалось доказать, что индуктивно-статистических вывод, использующий их, не приводит к противоречиям [7–8]. Для обнаружения таких максимально специфических правил был разработан специальный семантический вероятностный вывод, который можно рассматривать как метод обнаружения максимально специфических причинно-следственных связей. На основе этого вывода разработана формальная модель нейрона, удовлетворяющая правилу Хебба, в которой этот вывод обнаруживает максимально специфические причинно-следственные связи [9]. Причинно-следственные связи могут заикливаться, создавая неподвижные точки циклически взаимно предсказывающихся атрибутов. Эти неподвижные точки имеют особый смысл и отражают «естественную» классификацию объектов внешнего мира. В сознании они проявляются как перцептивные циклы. Хороший пример таких циклов приводит У. Найсер: "Мы записали на видеомагнитофон две "игры" (например, футбол и хоккей – Е.В.), а затем с помощью зеркала осуществили полное визуальное наложение двух передач – как если бы на телевизионном экране одновременно демонстрировались два канала... Испытуемых просили наблюдать за одной игрой и игнорировать другую, нажимая на ключ при каждом целевом событии (например, при каждом ударе по мячу, шайбе – Е.В.)... При темпе 40 целевых событий в минуту было одинаково легко следить за игрой независимо от того, демонстрировалась она вместе с другой или отдельно. Количество ошибок составляло

примерно 3%... Естественность этой задачи и отсутствие интерференции со стороны второго эпизода просто удивительны. Испытуемый *не видит* иррелевантную игру... Циклическая модель восприятия позволяет легко объяснить эти результаты" [10].

Строение объектов внешнего мира впервые было проанализировано в области «естественной» классификации. Было замечено, что «естественные» классы животных или растений отличаются потенциально бесконечным множеством свойств [11]. Естествоиспытатели, строившие «естественные» классификации, отмечали, что построение «естественной» классификации заключается в «индикации» – от бесконечно большого числа признаков нужно перейти к ограниченному их количеству, которое заменило бы все остальные признаки [12]. Это означает, что в «естественных» классах признаки сильно коррелированы, например, если есть 128 классов и атрибуты двоичные, то независимыми «индикаторными» атрибутами среди них будут около 7 атрибутов, т.к. $2^7 = 128$, а другие атрибуты могут быть предсказаны по значениям этих 7 атрибутов. Мы можем выбирать различные 7–15 атрибутов в качестве «индикаторных» и тогда другие атрибуты, которых потенциально бесконечное количество, могут быть предсказаны на основе этих выбранных атрибутов. Поэтому существует огромное множество причинно-следственных связей между атрибутами «естественных» классов.

Мы формализуем «естественную» классификацию путем обобщения анализа формальных понятий [13]. Формальные понятия могут быть определены как неподвижные точки детерминированных правил (не имеющих исключений) [13]. Мы обобщаем формальные понятия на вероятностный случай, заменяя детерминированные правила вероятностными максимально специфическими правилами и определяя вероятностные формальные понятия как неподвижные точки этих максимально специфических правил [14–15]. Можно показать [16], что вероятностные формальные понятия формализуют «естественную» классификацию, если их применить к некоторой выборке из генеральной совокупности. При этом полученная «естественная» классификация будет удовлетворять всем требованиям, которые естествоиспытатели предъявляли к «естественным» классификациям [16].

Алгоритм «естественной» классификации основан на определенном критерии максимальной согласованности причинных связей по взаимному предсказанию, который близок по смыслу интегрированной информации и тоже «измеряет» «интегрированность» причинных связей. Подробнее он рассмотрен в последнем разделе.

«Естественная» классификация и «естественные» понятия

Высокая корреляция признаков для «естественных» классов была подтверждена и в когнитивной науке. Eleanor Rosch сформулировала прин-

ципы категоризации, одним из которых является следующий: «воспринимаемый мир не является неструктурированным набором свойств, обнаруживаемых с равной вероятностью, напротив, объекты воспринимаемого мира имеют сильно коррелированную структуру» [17–18]. Непосредственно воспринимаемые объекты (basic objects) – информационно богатые связки наблюдаемых и функциональных свойств, которые образуют естественную разрывность, создающую категоризацию. Позже Bob Rehder предложил теорию причинных моделей, в которой отношение объекта к категории основывается уже не на наборе атрибутов, а на близости порождающих причинных механизмов: «объекты классифицируются как члены некоторой категории в той степени, в которой его свойства, вероятно, были сгенерированы причинными законами данной категории» [19]. Таким образом, структура причинно-следственных связей между атрибутами объектов берется за основу категоризации. Поэтому «естественные» объекты воспринимаются не как набор атрибутов, а как «резонирующая» (с точки зрения клеточных ансамблей нейронов) система причинно-следственных связей. В то же время «резонанс» возникает тогда и только тогда, когда эти причинно-следственные связи отражают некоторую целостность некоторого «естественного» класса, в котором потенциально бесконечное количество атрибутов взаимно предсказывают друг друга. Для формализации причинных моделей Bob Rehder предложил использовать причинно-следственные графические модели [20]. Однако эти модели основаны на «развертывании» байесовских сетей, которые не допускают циклов и не могут формализовать циклические причинно-следственные связи. Разработанные нами вероятностные формальные понятия непосредственно формализуют циклические причинно-следственные связи [14–16].

Нейрон передает свое возбуждение другим нейронам через множество возбуждающих и тормозных синапсов. Ингибирующие синапсы могут тормозить нейроны. Это важно для «подавления» восприятия альтернативных образов, атрибутов и свойств. В рамках нашей формальной модели это достигается путем обнаружения «тормозных» причинно-следственных связей, которые предсказывают отсутствие атрибута/свойства объекта (воспринимаемый объект не должен иметь соответствующий атрибут/свойство). В формальной модели это определяется путем введения отрицаний для предикатов соответствующих атрибутов/свойств. Нами доказано [15–16], что в неподвижных точках, использующих только максимально специфические причинно-следственные связи, не возникает противоречий – не возникает ситуации, когда одновременно предсказывается наличие некоторого атрибута/свойства и его отсутствие.

Следует особо отметить, что «резонанс» взаимных предсказаний вос-

принимаемых свойств объектов (стимулов) осуществляется непрерывно во времени и поэтому предсказанные свойства должны точно совпадать с тем, что воспринятыми. Отсутствие противоречий в предсказаниях также является отсутствием противоречий между предсказанными стимулами и воспринятыми стимулами.

Альтернативный способ измерения интегрированной информации

Если «естественная» классификация описывает объекты внешнего мира, а когнитивные науки – восприятие объектов внешнего мира, то теория интегрированной информации, с нашей точки зрения, анализирует информационные процессы мозга по восприятию «естественной» классификации объектов внешнего мира. В этом случае феноменологические свойства, вводимые G. Tononi, могут быть проинтерпретированы в терминах «естественной» классификации следующим образом:

1) composition – «естественные» классы в виде причинных циклов образуют иерархию «естественных» классов;

2) information – совокупность различий приводит к качественному отличию или, иначе, качественно отличающиеся объекты различны по целой совокупности различных свойств. В нашей формализации циклические причинные связи автоматически формируют паттерны различающихся свойств;

3) integration – значима только система «резонирующих» причинных связей, свидетельствующая об избытке информации и восприятии высоко коррелированной структуры «естественного» объекта);

4) exclusion – только значения признаков, максимально взаимосвязанных причинными связями формируют «образ» или «прототип».

Теоретические результаты о непротиворечивости индуктивно-статистического вывода и непротиворечивости неподвижных точек предполагают, что известна вероятностная мера событий. Однако если мы обнаруживаем причинно-следственные связи на обучающей выборке и предсказываем свойства нового объекта, случайно выбранного из генеральной совокупности, или распознаем новый объект как член некоторого «естественного» класса, то противоречия могут возникать.

В этом случае мы используем определенный критерий максимальной согласованности причинных связей по взаимному предсказанию, основанный на информационной мере, близкой по смыслу к энтропийной мере интегрированной информации [21]. Сравним эти критерии.

Интегрированная информация G. Tononi основывается на эффективной информации ei , генерируемой системой. Она определяется как разница энтропии априорного состояния системы и апостериорного состояния системы, приводящего к некоторому состоянию x_1 через причинные связи [2]:

$$ei(X_0 \rightarrow x_1) := H(p^{\max}(X_0)) - H(p(X_0 \rightarrow x_1)),$$

где $p^{\max}(X_0)$ – распределение на состояниях системы с максимальной энтропией (maxent), $p(X_0 \rightarrow x_1)$ – распределение на состояниях системы, которые приводят к состоянию x_1 через причинные связи, $H(p(-))$ – энтропия распределения p .

Эффективная информация показывает, сколько информации генерируется системой без учета ее интеграции. Для определения интегрированной информации надо знать, сколько информации генерируется системой как целым по отношению к информации, генерируемой независимо ее частями. Для этого определяется некоторое минимальное информационное разбиение P^{MP} и интегрированная информация приобретает вид [2]:

$$\phi(x_1) = H \left[p(X_0 \rightarrow x_1) \parallel \prod_{M^k \in P^{MP}} p(M_0^k \rightarrow \mu_1^k) \right].$$

Опираясь на интегрированную информацию, G. Tononi определяет комплексы (которые, с нашей точки зрения, способны улавливать комплексы стимулов объектов «естественных» классов) как систему S с входным состоянием s_1 , $\phi(s_1) > 0$, такую, что S не содержится в некотором большем подмножестве со строго большим значением интегрированной информации ϕ . Это означает, что в комплексе достигается локальный максимум интегрированной информации.

Аналогично, для определения «естественных» классов алгоритм «естественной» классификации использует локальный максимум критерия согласованности причинных связей по взаимному предсказанию свойств объектов. Поскольку, согласно теории, «естественные» классы образуют неподвижную точку взаимных предсказаний признаков класса, то в этой неподвижной точке должен достигаться максимум согласованности причинных связей по предсказанию и одновременно максимум информационной взаимосвязи признаков класса. Поэтому критерий согласованности имеет следующий вид [21]:

$$\text{Krit}(X) = \sum_{R \in S(X)} H(R) - \sum_{R \in F(X)} H(R), \text{ где } H(R) = -\log(1 - \text{prob}(R)),$$

где частная энтропия $H(R)$ учитывает не саму условную вероятность $\text{prob}(R)$ причинной связи R , а ее близость к 1. Множество причинных связей $S(X)$ – это те, которые подтверждаются в неподвижной точке (предсказываемое значение совпадает с существующим в неподвижной точке), и $F(X)$ – это те, которые опровергаются в неподвижной точке. Проведенные эксперименты [21] показывают, что данный критерий

успешно находит неподвижные точки даже при неполной информации об объектах.

Заключение

Приведенные результаты позволяют предположить, что на основе отражения мозгом максимально специфических причинно-следственных связей и возникающих неподвижных точек можно создать математически точную модель отражения мозгом реальности, проявляющуюся в сознании. Для этого достаточно уметь точно отражать причинные связи с помощью максимально специфических причинных связей. Можно показать, что отражение мозгом максимально специфических причинно-следственных связей позволяет моделировать множество когнитивных функций в соответствии с существующими физиологическими и психологическими теориями. Организация целенаправленного поведения моделируется причинно-следственными связями между действиями и их результатами [22], которые полностью соответствуют теории функциональных систем [23]. Неподвижные точки адекватно моделируют восприятие [21]. Набор причинно-следственных связей моделирует экспертные знания [24].

Список литературы

1. Tononi G. An information integration theory of consciousness // BMC 2004. Neurosci 5:42.
2. David Balduzzi, Giulio Tononi. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework // PLoS Computational Biology. 2008. 4(6). 2–18.
3. Ozumi M., Albantakis L., Tononi G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology. 2014 May. V. 10. Issue 5.
4. Carnap R. Philosophical Foundations of Physics. Basic Books. 1966.
5. Hempel C. Aspects of Scientific Explanation // Aspects of Scientific Explanation and other Essays in the Philosophy of Science. The Free Press: New York. 1965.
6. Hempel C. Maximal Specificity and Lawlikeness in Probabilistic Explanation // Philosophy of Science 1968, 35, 16–33.
7. Vityaev E. The logic of prediction // Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference, Novosibirsk, Russia, August 16-19, 2005. Goncharov, S., Downey, R., Ono, H., Eds. World Scientific: Singapore. 2006. P. 263-276.
8. Vityaev E., Odintsov S. How to predict consistently? // Trends in Mathematics and Computational Intelligence. Studies in Computational Intelligence. 2019. Vol. 796. P. 35-41.
9. Vityaev E. A formal model of neuron that provides consistent predictions // Biologically Inspired Cognitive Architectures 2012. Proceedings of the Third Annual Meeting of the BICA Society (Chella, A., et al. eds.). Advances in Intelligent Systems and Computing, V. 196, Springer. 2013. P. 339-344.

10. Найсер У. Познание и реальность. Москва: Прогресс. 1981, 229 с.
11. Mill J. *System of Logic, Ratiocinative and Inductive*. L. 1983.
12. Смирнов Е.С. Конструкция вида таксономической точки зрения // Зоол. Журн. 1938. Т. 17, № 3. С. 387–418.
13. Ganter B. *Formal Concept Analysis: Methods, and Applications in Computer Science*. TU: Dresden, Germany. 2003.
14. Vityaev E., Demin A., Ponomaryov D. Probabilistic Generalization of Formal Concepts // *Programming and Computer Software*. 2012. 38:5, 219–230.
15. Vityaev E., Martinovich V. Probabilistic Formal Concepts with Negation // *PCI 2014*, Voronkov A., Virbitskaite I., Eds., LNCS 8974. 2015. P. 385-399.
16. Витяев Е.Е., Мартынович В.В. Формализация "естественной" классификации и систематики через неподвижные точки предсказаний // *Сибирские электронные математические известия (Siberian Electronic Mathematical Reports)*, Том 12, Институт математики им. С. Л. Соболева СО РАН. 2015. С. 1006–1031.
17. Rosch E. Natural categories // *Cognitive Psychology*. 1973. 4, 328–350.
18. Rosch E. Principles of Categorization // *Cognition and Categorization*, Rosch, E., Lloyd B., eds.; Lawrence Erlbaum Associates, Publishers: Hillsdale. 1978. P. 27–48.
19. Rehder B. Categorization as causal reasoning // *Cognitive Science*. 2003. 27, 709–748.
20. Rehder Bob, Martin J. Towards A Generative Model of Causal Cycles // *33rd Annual Meeting of the Cognitive Science Society 2011, (CogSci 2011)*, Boston, Massachusetts, USA, 20-23 July 2011. V.1, P. 2944–2949.
21. Витяев Е.Е., Неупокоев Н.В. Формальная модель восприятия и образа как неподвижной точки предвосхищений // *Подходы к моделированию мышления. М.: УРСС Эдиториал*. 2014. С. 155–172.
22. Vityaev E. Purposefulness as a Principle of Brain Activity // *Anticipation: Learning from the Past 2015*. Ed. Nadin, M., *Cognitive Systems Monographs*, V.25, Chapter No.:13, Springer. P. 231–254.
23. Anokhin P.K.: *Biology and neurophysiology of the conditioned reflex and its role in adaptive behaviour* 1974, Oxford etc.: Pergamon press. P. 574.
24. Vityaev E., Perlovsky L., Kovalerchuk B., Speransky S. Probabilistic dynamic logic of cognition // *Biologically Inspired Cognitive Architectures 2013*. Special issue: Papers from the Fourth Annual Meeting of the BICA Society (BICA 2013). V. 6. P. 159–168.