УДК: 53.02+519.7+519.812.2

Сознание как логически непротиворечивая прогностическая модель реальности*

Е.Е. Витяев[1,2]

[1]Институт математики им. С. Л. Соболева СО РАН, vityaev@math.nsc.ru
[2]Новосибирский Государственный Университет

Аннотация. В работах П.К. Анохина показано, что принцип опережающего отражения действительности является основой всего живого. Оно основано на причинности внешнего мира. В работе показывается, что сознание есть отражение мозгом причинности внешнего мира, которое может быть представлено в виде логически непротиворечивой прогностической модели реальности, которая непрерывно во времени и в пространстве проверяет себя на адекватность этой реальности. Показывается связь этой модели со следующими теориями: восприятия, интегрированной информации, функциональных систем и других. Приводятся компьютерные эксперименты, демонстрирующие эффективность данной модели.

Ключевые слова: Сознание, восприятие, интегрированная информация

Consciousness as a logically consistent prognostic model of reality

E.E. Vityaev[1,2]

[1]Sobolev institute of mathematics SD RAS, vityaev@math.nsc.ru
[2]Novosibirsk state university

Annotation. In the works of P.K. Anokhin showed that the principle of anticipatory reflection of reality is the basis of all living things. It is based on the causality of the external world. This work shows that consciousness is a reflection of the causality of the external world by the brain, which can be presented in the form of a logically consistent predictive model of reality that continuously in time and in space tests itself for the adequacy of this reality. The connection of this model with the following theories is shown: perception, integrated information, functional systems, and others. Computer experiments demonstrating the effectiveness of this model are presented.

Keywords: Consciousness, perception, integrated information

1. СОЗНАНИЕ и ПРИЧИННОСТЬ

В работах П.К. Анохина показано, что принцип опережающего отражения действительности является основой всего живого. Опережающее отражение действительности основано на причинности внешнего мира. В данной работе показывается, что отражение мозгом причинности внешнего мира может быть организовано в ***логически непротиворечивую прогности-***

*Эта работа поддержана проектом РФФИ № 15-07-03410-а

ческую модель реальности, которая непрерывно во времени и в пространстве проверяет себя на адекватность реальности. Эта модель и есть ***сознание***. При этом сознание отражает не все причинные связи, а только те, которые связаны с достижением организмом определенных целей (удовлетворением потребностей). В соответствии с информационной теорией эмоций П.В. Симонова, субъективные эмоциональные ощущения возникают только тогда, когда есть вероятностный прогноз достижения цели(ей) и не удовлетворенная потребность, соответствующая этим целям. Поэтому сознание всегда окрашено некоторыми субъективными ощущениями (qualia), которые отражают достигаемые организмом цели вместе с оценки вероятности достижения этих целей. При этом автоматизмы, не содержащие вероятностные прогнозы достижения целей (промежуточных целей), уходят из сознания.

Хороший пример целевой направленности сознания приводит У. Найсером (1981: 103-105): "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.) в наблюдаемой игре. ... 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*** (выделение – Е.В.) иррелевантную игру ... Циклическая модель восприятия позволяет легко объяснить эти результаты".

Предлагаемая нами в работе оригинальная формализация причинных связей и циклических причинных связей включает в себя циклическую модель восприятия (Витяев Е.Е. Неупокоев Н.В. 2014). Эта циклическая модель непрерывно во времени и в пространстве проверяет себя на адекватность реальности, как это описано в книге С.Д. Смирнова (1985: 153) «Психология образа»: «Все это позволяет нарисовать следующую картину хода познавательной деятельности на уровне восприятия. Индивид всегда имеет некоторый образ или модель окружения, которая непрерывна во времени и пространстве и носит прогностический характер, т.е. в ней экстраполируются и воспроизводятся на языке чувственных модальностей ожидаемые результаты воздействия источника стимула на наши органы чувств».

Сознание, как логически непротиворечивая прогностическая модель реальности, не только непрерывно во времени и пространстве прогнозирует, какие стимулы будут восприняты в следующий момент, но и непрерывно проверяет правильность этих прогнозов. Совпадение их с реально поступающими стимулами создает ощущение присутствия во внешнем мире.

В нашей формализации причинные связи обнаруживаются нейронами, формальная модель которых (Vityaev E.E. 2013) удовлетворяет правилу Хебба (1949). Циклические причинные связи обнаруживаются клеточными ансамблями, которые, как мы покажем, обнаруживают «естественную» классификацию объектов внешнего мира (Витяев Е.Е. Мартынович В.В. 2015) и обладают свойством интегрированной информации по G.Tononi (M. Oizumi L. Albantakis G. Tononi 2014).

Важным свойством сознания, которое проявляется не во всех структурах мозга, как это показано у G.Tononi (M. Oizumi L. Albantakis G. Tononi 2014), является его интеграционный характер. Наиболее емкой фразой, характеризующей этот характер, является: «различия делающие различие» ("differences that make a difference"). Она означает, что совокупность различий приводит к качественному отличию или иначе качественно отличающиеся объекты различны по целой совокупности различных свойств. Что бы уловить это свойство сознания G.Tononi вводит понятие интегрированной информации, когда информация от системы причинных связей свойств объекта превосходит информацию от совокупности свойств самих по себе. Тогда объекты различаются не по отдельным свойствам, а по совокупности (паттерну) различающихся свойств. В нашей формализации циклические причинные связи автоматически формируют паттерны различающихся свойств, которые причинно взаимосвязаны. В нашем эксперименте на закодированных цифрах, приводимом ниже, рассматриваются причин-

ные связи между различными признаками цифр, которые *верны для всех цифр*, но при этом выделение и идентификация *отдельных цифр* происходит только по паттернам свойств цифр, циклически взаимно предсказывающих друг-друга причинными связями. Более того, сами паттерны свойств в нашей формализации автоматически формируются причинными связями, как циклически взаимно предсказывающийся набор стимулов, формирующий неподвижную точку взаимопредсказаний. Не все области мозга способны формировать богатые циклические причинные связи для идентификации объектов внешнего мира, а также для непрерывной во времени и пространстве проверки правильности сделанных предсказаний, что создает ощущение реального существования объектов во внешнем мире.

Свойство сознания по восприятию объектов паттернами свойств, в которых «различия делают различие», на самом деле опирается на такое свойство внешнего мира, как высокая коррелированность свойств «естественных» объектов внешнего мира. Это свойство подтверждено естествоиспытателями, которые строили «естественные» классификации, а также исследованиями по формированию «естественных» понятий в когнитивных науках. Например, для формализации «естественных» понятий Bob Rehder (2003) выдвинул теорию причинных моделей, в которой отношение объекта к категории основывается уже не на множестве признаков и близости по признакам, а на основании сходства порождающего причинного механизма: «объект классифицируется как член некоторой категории в той степени, в которой его свойства, вероятно, были сгенерированы причинными законами данной категории». Более подробно эти исследования рассмотрены в разделах 4-5.

Сознание, как и «естественная» классификация объектов внешнего мира иерархична. Как говорит Дж. Гибсон (1988), внешний мир имеет свойство «встроенности»: «… ущелья встроены в горы, деревья встроены в ущелья, листья встроены в деревья, клетки встроены в листья. При любом масштабе можно обнаружить, что одни формы содержат в себе другие». На каждом уровне детальности формируются свои «естественные» классы и образы по одним и тем же законам. Наиболее общим образом является «образ мира», который формируется, начиная с первых дней жизни. Субъект никогда не воспринимает отдельные объекты, а всегда воспринимается некоторая целостная картина. Отдельные объекты выделяются в «видимом поле» (Дж. Гибсон 1988, Столин В.В. 1976) в процессе деятельности с этими объектами как с предметами. Предметная деятельность преобразует «видимое поле» в «видимый мир» и «образ мира» (Леонтьев А.Н. 1983). Это не тривиальный процесс и требует многослойной нейронной обработки, как показано в экспериментах с Deep Learning. Наша формализация в виде циклических причинных связей может осуществлять такую же иерархическую обработку видимого поля в виде иерархии «естественных» классов в соответствии со встроенностью объектов реальности и, как показано в наших экспериментах (Витяев Е.Е. Неупокоев Н.В. 2012), делает это принципиально точнее, чем Deep Learning.

Восприятие отдельного объекта начинается не с объекта, а с «образа мира» (Смирнов С.Д. 1985: 144): «Образ мира не складывается из образов отдельных явлений и предметов, а с самого начала развивается и функционирует как некоторое целое. Это значит, что любой образ есть не что иное, как элемент образа мира, и сущность его не в нем самом, а в том месте, в той функции, которую он выполняет в целостном отражении реальности. Эта характеристика образа мира определяется взаимосвязями и взаимозависимостями между элементами самой объективной реальности. ...". С информационной точки зрения, чем выше иерархия некоторого «естественного» класса, тем он устойчивей и инвариантен по отношению к всевозможным вариациям соответствующего образа. Как неподвижная точка взаимопредсказаний свойств, более высокие классы описываются причинными связями с более высокими оценками вероятности. В этом случае более высокие и инвариантные классы начинают доминировать над нижележащими классами и играть ведущую роль.

У. Найсер (1981: 66-67) отмечал, что восприятие как процесс не завершено, пока перцептивный цикл не завершится: «Тахистоскопические эксперименты попросту не относятся к нормальным перцептивным навыкам, и термин «восприятие» в полном смысле слова нельзя отнести к тому, что в них происходит. Такая интерпретация позволяет объяснить, почему ин-

троспективные отчеты в тахистоскопических исследованиях столь противоречивы». Эксперименты с маскировкой стимулов также показывают, что пока причинные связи не зациклились и не привели к устойчивому циклическому возбуждению (неподвижной точке) предвосхищений и проверке совпадения их с реальными стимулами, мы не можем осознать эту реальность. Одна из ролей сознания – ликвидация противоречий в осознании поступающих стимулов. С информационной точки зрения это старая и до сих пор не решенная проблема философии науки – проблема статистической двусмысленности (см. следующий раздел). При обнаружении причинных связей на данных, можно обнаружить правила, которые будут приводить к противоречиям. В нашей формализации причинные связи предсказывают не только наличие свойства (стимула), но и его отсутствие. Таким образом, моделируется не только возбуждение, но и торможение в нейронной сети. Для разрешения противоречий К.Гемпель предложил использовать максимально специфические правила. Нами эта проблема решена и предложена формализация максимально специфических правил, для которых не возникает противоречий (E. Vityaev 2006). Предложена также формальная модель нейрона, удовлетворяющая правилу Хебба и обнаруживающая причинные связи как замыканий условных связей на уровне нейрона, обнаруживающая (в пределе) максимально специфические причинные (условные) связи. Но на этом проблемы не заканчиваются – для формализации неподвижных точек предсказания также необходимо было доказать, что в них не возникает противоречий. Это осуществлено в работах, где неподвижные точки предсказаний формализованы в виде вероятностных формальных понятий (Vityaev Demin Ponomaryov 2012, Vityaev Martinovich 2014). Такие неподвижные точки взаимно предсказывают не только наличие стимулов, но и их отсутствие, которое осуществляет вытормаживание не совместимых с данными предсказаниями стимулов. В частности, это моделирует восприятие иллюзий типа фигура-фон, когда есть два не совместимых образа в виде двух различных неподвижных точек. Для осознания в целом поступающей стимуляции, нужно разрешить противоречия между возникающими неподвижными точками, входящими в стимуляцию образов. В разрешении противоречий главную роль играют наиболее сильные неподвижные точки наиболее инвариантных классов. В любом случае, пока не сформируется непротиворечивое согласование всех неподвижных точек воспринимаемых в данный момент образов, включая «образ мира» формируя целостную картину воспринимаемой реальности, осознание воспринятого не возникает.

Прогнозирование стимулов и сличение их с реальностью осуществляется не только в клеточных ансамблях Хебба, но и через деятельность, когда предвосхищение некоторого стимула в результате некоторого (перцептивного) действия по внутреннему контуру мозга (см. рис. 5) сопоставляется с реально поступающими стимулами, полученными в результате осуществления этого действия по внешнему контуру. Такое предвосхищение включает процесс прогнозирования достижения целей, как описано в теории функциональных систем и изложено далее в разделах 7-8.

2. Причинность и принцип опережающего отражения действительности

Причинность является следствием *физического детерминизма*: «для всякой изолированной физической системы, произвольно фиксированное состояние системы детерминируют все последующие состояния» (Закон 1967). Но возьмем, например, автомобильную аварию (Карнап 1971), что явилось её причиной? Это может быть состояние поверхности дороги, её влажность, расположение солнца относительно взоров водителей, нарушение правил дорожного движения, психологическое состояние водителей, исправность тормозов и т.д. Понятно, что в этом случае нет определенной причины происшествия. Понятие причинности анализировал Д.Юм, но, как правильно отмечается в (Редько 2011) он «не нашел никакого другого основания, кроме некоторого внутреннего чувства привычки». Причинность не только улавливается мозгом, но и сыграла важную роль в процессе когнитивной эволюции в формировании информационных процессов работы мозга (Редько 2011).

В философии науки причинность сводится к предсказанию и объяснению. «Причинное отношение означает предсказуемость … в том смысле, что, если полная предыдущая ситуация будет известна, событие может быть предсказано …, если будут даны все относящиеся к событию факты и законы природы» (Карнап 1971). Понятно, что всех фактов, число которых в случае аварии потенциально бесконечно, и всех законов никто знать не может. Некоторые из законов могут быть обнаружены на данных. Поэтому причинность сводится к предсказанию в соответствии с индуктивно-номологическим выводом, состоящим в логическом выводе предсказаний из фактов и вероятностных законов с некоторой вероятностной оценкой.

При обнаружении законов (закономерностей) на реальных данных возникает проблема статистической двусмысленности, которая состоит в том, что в процессе обучения (индуктивного вывода) мы можем получать вероятностные правила, из которых выводится противоречие. Пример: наблюдая людей, можно вывести два правила: если человек философ, то он не миллионер, а если он держатель приисков, то он миллионер. Применяя эти два правила к известному философу П. Суппесу, мы получим противоречие: поскольку он философ, то он не должен быть миллионером, а, поскольку он держатель приисков, то должен быть миллионером.

Чтобы избежать противоречий Гемпель (1968) ввел требование максимальной специфичности для законов, состоящее в том, что закон должен учитывать максимум информации, относящейся к предсказываемому свойству. В нашем примере максимально специфичными должны быть правила: (1) если человек философ, но не держатель приисков, то он с ещё большей вероятностью не миллионер, а (2) если он держатель приисков, но не философ, то он также, с ещё большей вероятностью, миллионер. Применение этих двух правил уже не приводит к противоречиям, поскольку они не применимы к одним и тем же объектам.

Нами разработан специальный семантический вероятностный вывод (Vityaev 2006), который выводит максимально специфические правила. Таким образом, решается проблема статистической двусмысленности и получается подходящая формализация причинности, которая позволила получить нужные математические результаты:

1. доказано, что семантический вероятностный вывод обнаруживает причинные связи в виде максимально специфических правил, которые учитывают всю доступную информацию и предсказывают без противоречий, что решает проблему статистической двусмысленности (Vityaev 2006);
2. доказано, что неподвижные точки по максимально специфическим правилам непротиворечивы (Vityaev Martinovich 2014);
3. доказано, что неподвижные точки по максимально специфическим правилам являются вероятностным обобщением формальных понятий (Vityaev Martinovich 2014), исследуемых в анализе формальных понятий;
4. семантический вероятностный вывод может быть рассмотрен как формальная модель нейрона, а неподвижные точки как клеточные ансамбли (Vityaev 2013, 2015).

3. Причинность и формальная модель нейрона

С нашей точки зрения смысл деятельности нейронов состоит в обнаружении причинных связей. Приведем формальную модель нейрона, обнаруживающую максимально специфичные условные связи (Vityaev et al 2013).

Под *информацией* поступающей на «вход» мозга будем понимать всю воспринимаемую мозгом стимуляцию: мотивационную, обстановочную, пусковую, санкционирующую, обратную афферентацию о произведенных действиях, поступающую по коллатералям на «вход» и т. д. Из экологической теории восприятия Дж. Гибсона (1988) следует, что под информацией можно понимать любую характеристику энергетического потока света, звука и т.д., поступающую на «вход» мозга.

Определим информацию, передаваемую возбуждением некоторого нервного волокна на синапсы нейрона, одноместными предикатами $P_j^i(\mathsf{a}) \Leftrightarrow (x_i(\mathsf{a}) = x_{ij})$, $i = 1,...,n;\ j = 1,...,n_i$, где

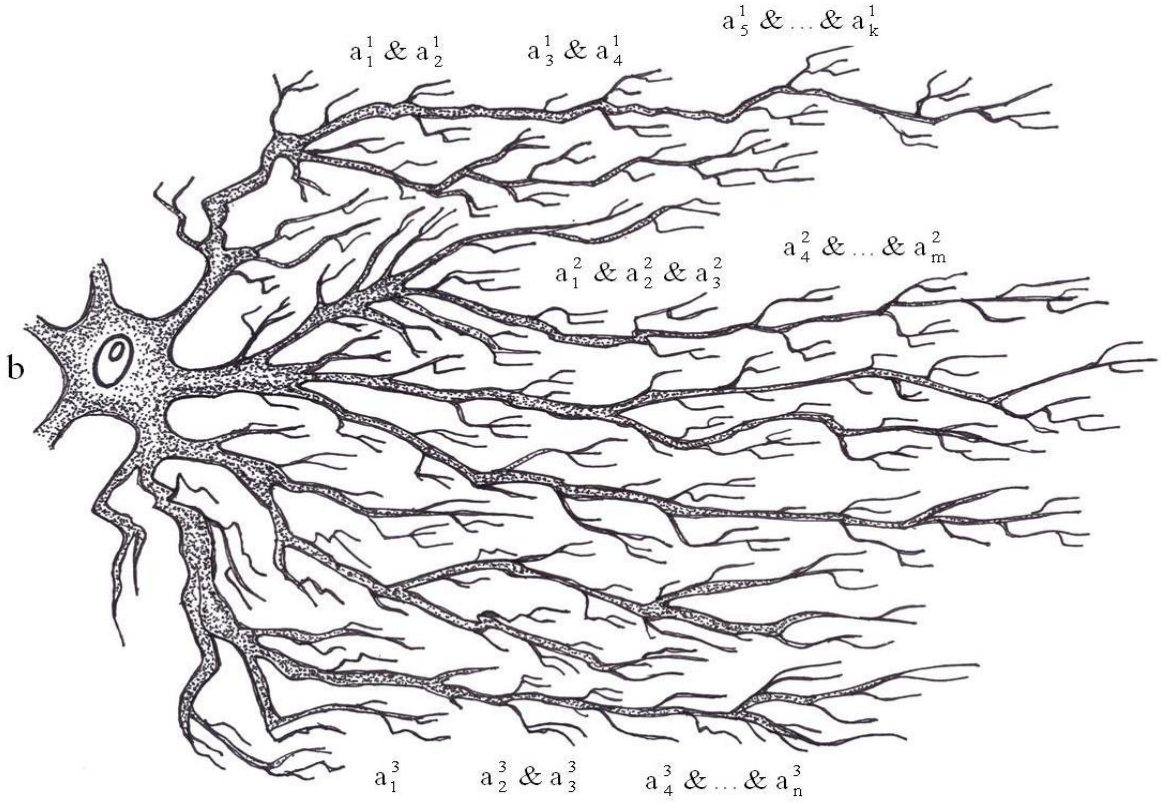

Рис. 1. Формальная модель нейрона.

$x_i(\mathsf{a})$ – информация, а x_{ij} – её значение в текущей ситуации (на объекте) a. Если информация передается на возбуждающий синапс, то она воспринимается нейроном, как истинность предиката $P_j^i(\mathsf{a})$, если на тормозной синапс, то, как ложность $\neg P_j^i(\mathsf{a})$ предиката. Возбуждение нейрона в ситуации a и передачу этого возбуждения на аксон нейрона определим предикатом $P_0(\mathsf{a})$. Если нейрон тормозится в ситуации a, то определим эту ситуацию как прогнозирование отрицания предиката $\neg P_0(\mathsf{a})$. Предикаты $P_j^i(\mathsf{a})$, $P_0(\mathsf{a})$ и их отрицания $\neg P_j^i(\mathsf{a})$, $\neg P_0(\mathsf{a})$ являются литералами (атомарными высказываниями или их отрицаниями), которые будем обозначать как $a,b,c,\ldots \in L$, где L - множество всех литералов в словаре $\{P_j^i\}, i=1,\ldots,n;\ j=1,\ldots,n_i$.

Каждый нейрон имеет свое рецептивное поле, возбуждающее его безусловно. Первоначальной (до всякого обучения) семантикой предиката P_0 является это рецептивное поле. В процессе обучения эта информация обогащается и может дать достаточно специализированный нейрон типа нейрона Билла Клинтона.

Мы предполагаем, что формирование условных связей на уровне нейрона происходит по правилу Хебба (1949). Мы формализуем правило Хебба с помощью семантического вероятностного вывода, который принципиально отличается от других формализаций тем, что обнаруживает максимально специфические условные связи.

В процессе семантического вероятностного вывода нейрон обнаруживает множество {R} правил (условных связей) вида:

$$R = (a_1 \& \ldots \& a_k \Rightarrow b),\ a_1,\ldots,a_k,b \in L, \qquad (1)$$

где $a_1,\ldots,a_k$ – предикаты для информации, приходящей на синапсы дендритов нейрона, а b – предикат для информации $P_0(\mathsf{a})$ или $\neg P_0(\mathsf{a})$ аксона нейрона.

Правила характеризуются оценкой условной вероятности, которая вычисляется следующим образом. Подсчитаем число случаев $n(a_1,...,a_k,b)$, когда произошло событие $\langle a_1,...,a_k,b\rangle$ – одновременное возбуждение/торможение входов $\langle a_1,...,a_k\rangle$ нейрона и самого нейрона непосредственно перед действием подкрепления (которое может быть, как положительным, так и отрицательным). Среди случаев $n(a_1,...,a_k,b)$ подсчитаем случаи $n^+(a_1,...,a_k,b)/n^-(a_1,...,a_k,b)$, когда подкрепление было положительным/отрицательным. Тогда оценка условной вероятности правила (1) равна:

$$\mu(b/a_1,...,a_k)=\frac{n^+(a_1,...,a_k,b)-n^-(a_1,...,a_k,b)}{n(a_1,...,a_k,b)}.$$

Если эта вероятность становится отрицательной, то это означает торможение нейрона с вероятностью, взятой с обратным знаком.

Приведем формальное определение семантического вероятностного вывода и формальной модели нейрона. Свойства этой модели приведены в конце раздела.

Под данными обучения *Data* будем понимать все случаи возбуждения или торможения нейрона, когда было подкрепление. Множество всех правил вида (1) обозначим через Pr.

Правило $R_1=(a_1^1\,\&\,a_2^1\,\&...\&\,a_{k_1}^1\Rightarrow c)$ будем называть *более общим*, чем правило $R_2=(b_1^2\,\&\,b_2^2\,\&...\&\,b_{k_2}^2\Rightarrow c)$, обозначим это как $R_1\succ R_2$, тогда и только тогда, когда $\{a_1^1,a_2^1,...,a_{k_1}^1\}\subset\{b_1^2,b_2^2,...,b_{k_2}^2\}$, $k_1<k_2$ и *не менее общим* $R_1\succsim R_2$, если $k_1\le k_2$.

Нетрудно доказать, что $R_1\succsim R_2\Rightarrow R_1\vdash R_2$ и $R_1\succ R_2\Rightarrow R_1\vdash R_2$, где $\vdash$ – доказуемость в исчислении высказываний. Таким образом, не менее общие (и более общие) высказывания логически сильнее. Кроме того, более общие правила проще, так как содержат меньшее число литер в посылке правила, поэтому отношение $\succ$ можно воспринимать как *отношение простоты* в смысле.

Определим множество предложений *F*, как множество высказываний, полученных из литер *L* замыканием относительно логических операций $\wedge,\vee,\neg$. *Вероятность* на множестве предложений *F* определим как отображение $\mu:F\mapsto[0,1]$, удовлетворяющее условиям:

1. Если $\vdash\varphi$, то $\mu(\varphi)=1$;
2. Если $\vdash\neg(\varphi\wedge\psi)$, то $\mu(\varphi\vee\psi)=\mu(\varphi)+\mu(\psi)$.

Определим условную вероятность правила $R=(a_1\,\&...\&\,a_k\Rightarrow c)$ как

$$\mu(R)=\mu(c/a_1\,\&...\&\,a_k)=\frac{\mu(a_1\,\&...\&\,a_k\,\&\,c)}{\mu(a_1\,\&...\&\,a_k)},\text{ если }\mu(a_1\,\&...\&\,a_k)>0.$$

Мы предполагаем, что оценка вероятности, определённая выше, в пределе дает вероятность μ. Множество всех правил из Pr, для которых условная вероятность определена, обозначим через Pr_0. *Вероятностным законом* будем называть такое правило $R\in\mathrm{Pr}_0$, которое нельзя обобщить (логически усилить), не уменьшив его условную вероятность, т.е. для любого $R'\in\mathrm{Pr}_0$, если $R'\succ R$, то $\mu(R')<\mu(R)$.

Вероятностные законы – это наиболее общие, простые и логически сильные правила, среди правил, имеющих не более высокую условную вероятность. Обозначим множество всех вероятностных законов через *PL*.

Формальную модель нейрона определим как множество всех вероятностных законов $\Phi=\{R\}, R\in PL$, которые обнаруживает нейрон. Отношение *вероятностного вывода* $R_1\sqsubseteq R_2$, $R_1,R_2\in PL$ определим как одновременное выполнение двух неравенств $R_1\succsim R_2$ и $\mu(R_1)\le\mu(R_2)$. Если оба неравенства строгие, то отношение вероятностного вывода будем называть строгим *отношением вероятностного вывода*

$R_1\sqsubset R_2\Leftrightarrow R_1\succ R_2\,\&\,\mu(R_1)<\mu(R_2)$.

Семантическим вероятностным выводом (Vityaev 2006) будем называть максимальную (которую нельзя продолжить) последовательность вероятностных законов, находящихся в отношении строгого вероятностного вывода $R_1 \sqsubset R_2 \sqsubset ... \sqsubset R_k$. Последний вероятностный закон R_k в этом выводе будем называть *максимально специфическим*.

Теорема [доказана в (Vityaev 2006)]**.** Предсказание по максимально специфическим правилам непротиворечиво.

Совокупность $\Phi = \{R\}, R \in PL$ всех вероятностных законов, всех семантических вероятностных выводов, которые обнаруживает нейрон в *процессе* обучения, составляет его формальную модель. В соответствии с этой моделью, нейрон $P_0(a)$ возбуждается/тормозится в соответствии с тем вероятностным законом его формальной модели, который применим в данной ситуации (посылка которого истинна) и имеет максимальную вероятность.

Опишем свойства полученной формальной модели нейрона:

1) нейрон осуществляет «замыкание условных связей». При обнаружении условных стимулов, позволяющих предсказывать с некоторой вероятностью возбуждение нейрона, образуется условная связь в виде правила (1). При обнаружении новых стимулов, позволяющих предсказывать возбуждение нейрона с ещё большей вероятностью, они присоединяются к данной условной связи;
2) возбуждение или торможение нейрона осуществляется по максимально вероятным правилам. Это подтверждается тем, что в процессе выработки условных связей, а также при замыкании условных связей на уровне нейрона, скорость ответа нейрона на условный сигнал, тем выше, чем выше вероятность условной связи. Поскольку максимально специфические правила, учитывающие всю имеющуюся информацию, одновременно являются максимально вероятными, то предсказание (возбуждение нейрона) осуществляется по этим правилам;
3) предсказание по максимально специфическим правилам, осуществляемое нейроном, в пределе непротиворечиво. Поэтому в процессе дифференциации условных связей нейрон обучается предсказывать без противоречий – срабатывают либо его возбуждающие максимально специфические правила, либо тормозные, но не одновременно;
4) на рис. 1 показано несколько семантических вероятностных выводов, осуществляемых нейроном. Например, условная связь $(b \Leftarrow a_1^1 \& a_2^1)$ усиливается новыми стимулами $a_3^1 \& a_4^1$ до связи $(b \Leftarrow a_1^1 \& a_2^1 \& a_3^1 \& a_4^1)$, если стимулы $a_3^1 \& a_4^1$ увеличивают условную вероятность предсказания возбуждения нейрона b.

   a) $(b \Leftarrow a_1^1 \& a_2^1) \sqsubset (b \Leftarrow a_1^1 \& a_2^1 \& a_3^1 \& a_4^1) \sqsubset (b \Leftarrow a_1^1 \& a_2^1 \& a_3^1 \& a_4^1 \& a_5^1 \& ... \& a_k^1)$;

   b) $(b \Leftarrow a_1^2 \& a_2^2 \& a_3^2) \sqsubset (b \Leftarrow a_1^2 \& a_2^2 \& a_3^2 \& a_4^2 \& ... \& a_m^2)$;

   c) $(b \Leftarrow a_1^3) \sqsubset (b \Leftarrow a_1^3 \& a_2^3 \& a_3^3) \sqsubset (b \Leftarrow a_1^3 \& a_2^3 \& a_3^3 \& a_4^3 \& ... \& a_n^3)$.

Семантический вероятностный вывод и реализующая его программная система Discovery успешно применялись для решения ряда прикладных задач.

4. Взаимосвязь «естественной» классификации, «естественных» классов и сознания как интегрированной информации.

Строение объектов внешнего мира впервые было проанализировано в области «естественной» классификации. Было замечено, что «естественные» классы животных или растений обладают и отличаются потенциально бесконечным множеством свойств (Mill 1843). 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 (Смирнов 1938, Забродин 1981). Это означает, что в «естественных» классах признаки сильно коррелированны, например, если классов 128 и признаки бинарные, то незави-

симыми «индикаторными» признаками среди них будут только 7 признаков, т.к. 2^7=128. Остальные признаками можно предсказать по закономерностям, связывающим эти 7 признаков с остальными признаками.

Высокая коррелированность признаков для «естественных» классов была подтверждена в когнитивных исследованиях. Eleanor Rosch сформулировала принципы категоризации, один из которых гласит: «воспринимаемый мир не является неструктурированным множеством равновероятно встречающихся свойств, наоборот, объекты воспринимаемого мира имеют высоко коррелированную структуру» (Rosch 1978). Непосредственно воспринимаемые объекты (basic objects) – информационно богатые связки наблюдаемых и функциональных свойств, которые образуют естественную разрывность, создающую категоризацию. В дальнейшем теория «естественных» понятий Eleanor Rosch получила название прототипической теории понятий (prototype theory). Основные ее черты описываются следующим образом: «The prototype view (or probabilistic view) keeps the attractive assumption that there is some underlying common set of features for category members but relaxes the requirement that every member have all the features. Instead, it assumes there is a probabilistic matching process: Members of the category have more features, perhaps weighted by importance, in common with the prototype of this category than with prototypes of other categories» (Ross et al 2008).

В дальнейших исследованиях было обнаружено, что моделей, основанных на признаках, сходстве и прототипах, недостаточно для описания классов. Необходимо учитывать теоретические, причинные и онтологические знания, относящиеся к объектам классов. Например, люди не только знают, что птицы имеют крылья, могут летать и вить гнезда на деревьях, но также и то, что птицы вьют гнезда на деревьях, потому что могут летать, и летать, потому что они имеют крылья.

В дальнейшем, Bob Rehder выдвинул теорию причинных моделей, в которой отношение объекта к категории основывается уже не на множестве признаков и близости по признакам, а на основании сходства порождающего причинного механизма: «объект классифицируется как член некоторой категории в той степени, в которой его свойства, вероятно, были сгенерированы причинными законами данной категории» (Rehder 2003). Таким образом, за основу классификации берется уже структура причинных зависимостей между признаками объектов. Для формализации причинных моделей Bob Rehder предложил causal graphical models (CGMs). Однако эти модели основываются на «развертывании» байесовских сетей, которые не допускают циклов (Rehder et al 2011).

Основная гипотеза данной работы: информационные процессы работы мозга и сознание настроилась в процессе эволюции на извлечение высоко коррелированной структуры признаков «естественных» объектов путем формирования «естественных» понятий объектов.

Это подтверждается работами G.Tononi (см. следующий раздел), который определяет сознание как проявление интегрированной информации – информации, генерируемой системой через причинное взаимодействие между ее частями (Oizumi Albantakis Tononi 2014). Однако он рассматривает интегрированную информацию как внутреннее свойство системы. В отличии от G.Tononi, мы рассматриваем интегрированную информацию как способность системы отражать высокую коррелированность признаков объектов «естественных» классов, а сознание, как способность комплексного иерархического отражения «естественной» классификации объектов внешнего мира.

5. Сознание как интегрированная информация.

Если «естественная» классификация описывает объекты внешнего мира, а когнитивные науки – восприятие объектов внешнего мира, то теория интегрированной информации анализирует информационные процессы мозга по восприятии объектов внешнего мира.

G.Tononi (2014) определяет сознание как первичное понятие, которое обладает следующими феноменологическими свойствами: composition, information, integration, exclusion. Для более точного определения этих свойств G.Tononi (2014) вводит понятие интегрированной

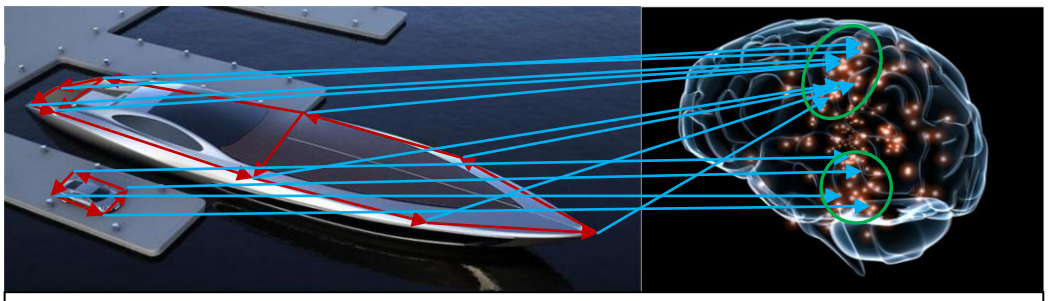

Рис. 2. Отражение мозгом причинных связей между свойствами объектов.

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. В скобках мы приводим интерпретацию этих свойств с точки зрения «естественной» классификации.

1. composition – elementary mechanisms (causal interactions) can be combined into higher-order ones («естественные» классы в виде причинных циклов образуют иерархию «естественных» классов) ;
2. information – only mechanisms that specify ‘differences that make a difference’ within a system count (объяснено выше);
3. integration – only information irreducible to non-interdependent components counts (значима только система «резонирующих» причинных связей, свидетельствующая об избытке информации и восприятии высоко коррелированной структуры «естественного» объекта);
4. exclusion – only maxima of integrated information count (только значения признаков, максимально взаимосвязанных причинными связями формируют «образ» или «прототип»).

Поскольку G.Tononi не рассматривает «естественную» классификацию объектов внешнего мира, то эти свойства определяются как внутренние свойства системы. Мы рассмотрим эти свойства не как внутренние свойства системы, а как способность системы отражать комплексы причинных связей внешних объектов, а сознание – как способность комплексного иерархического отражения «естественной» классификации внешнего мира.

Рассмотрим процесс отражения причинных связей (рис. 2). 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 структур мозга в системы, обозначенные овалами.

В теории G.Tononi рассмотрен только третий пункт процесса отражения. Квалиа (quale) G.Tononi определяет как всю совокупность возбужденных групп нейронов с максимально интегрированной концептуальной структурой (maximally integrated conceptual structure).

Интегрированная информация у G.Tononi рассматривается как система циклических причинных связей. Однако остается непонятным, что же отражает интегрированная информация. Нами выдвигается гипотеза о том, что «естественная» классификация, «естественные» понятия и интегрированная информация G.Tononi описываются одним и тем же формализмом и в определенном смысле тождественны друг другу. Мозг с помощью интегрированной информации настраивается на восприятие «естественных» объектов внешнего мира.

6. Единая формализация «естественной» классификации, «естественных» понятий и сознания, как интегрированной информации по G.Tononi

Нами предлагается принципиально новый математический аппарат для определения интегрированной информации, «естественной» классификации и «естественных» понятий, основанный на прямой формализации «резонанса» причинных связей. Мы предполагаем, что мозг осуществляет все возможные выводы по причинным связям, отражая высоко коррели-

рованную структуру объектов внешнего мира. Эти причинные связи в процессе восприятия «естественных» объектов замыкаются на себя, образуя определенный «резонанс», который является системой с высоко интегрированной информацией в смысле G.Tononi. При этом, «резонанс» возникает тогда и только тогда, когда эти причинные связи отражают некоторый целостный «естественный» объект, в котором потенциально бесконечное множество признаков взаимно предполагают друг-друга. Возникающие при этом циклы выводов по причинным связям математически описываются «неподвижными точками», которые характеризуются тем, что дальнейшее применение выводов к рассматриваемым свойствам не предсказывает новых свойств. Полученное в неподвижной точке множество взаимно предсказывающихся свойств дает «образ» класса и «прототип» объекта. Поэтому мозг воспринимает «естественный» объект не набором признаков, а как «резонирующую» систему причинных связей замыкающихся на себя через одновременный вывод всей совокупности признаков «образа» или «прототипа» образующих целостный паттерн.

Рассмотрим пример компьютерного моделирования обнаружения «естественных» классов, «естественных» понятий и интегрированной информации для закодированных цифр.

Пусть $X(a)$ – множество свойств объекта *a*, заданных некоторым множеством предикатов, а $(P_{i_1} \& ... \& P_{i_k} \Rightarrow P_{i_0}) \in MS(X)$ – множество максимально специфических условных связей (см. раздел 3), выполненных для свойств X, $\{P_{i_1},...,P_{i_k}\} \subset X$. Тогда оператор предсказания Pr и неподвижная точка могут быть записаны следующим образом (Витяев Неупокоев 2012, 2014):

$$Pr(X) = \Phi_{Krit}(X \cup \{P_{i_0} \mid (P_{i_1} \& ... \& P_{i_k} \Rightarrow P_{i_0}) \in MS(X)\} \cup \{\neg P_{i_0} \mid (P_{i_1} \& ... \& P_{i_k} \Rightarrow \neg P_{i_0}) \in MS(X)\}),$$

где $\Phi_{Krit}(X)$ – оператор, модифицирующий множество признаков X путем добавления или удаления некоторого признака так, чтобы определенный критерий Krit взаимной согласованности причинных связей по взаимному предсказанию признаков из X был максимальным. Критерий Krit иначе измеряет информационную интеграцию признаков по системе причинных связей $MS(X)$, чем это делается в теории G.Tononi. Неподвижная точка достигается тогда, когда $Pr^{n+1}(X(a)) = Pr^n(X(a))$, для некоторого n, где Pr^n – n кратное применение оператора Pr. Поскольку при каждом применении оператора Pr значение критерия Krit увеличивается и в неподвижной точке достигает локального максимума, то неподвижная точка, отражающая некоторый «естественный» объект, обладает максимумом интегрированной информации и свойством «exclusion» по G.Tononi. Однако принципиальным отличием данного подхода от теории G.Tononi является то, что система причинных связей в неподвижной точке сама, без внешних по отношению к ней оценок, формирует набор признаков «образа» или «прототипа» при восприятии «естественного» класса.

Проиллюстрируем формирование неподвижных точек компьютерным экспериментом по обнаружению «естественных» классов/понятий закодированных цифр.

Закодируем цифры как показано на рис. 3. Сформируем обучающее множество, состоящее из 360 перетасованных цифр (12 цифр рис. 3 продублированных в 30-ти экземплярах без

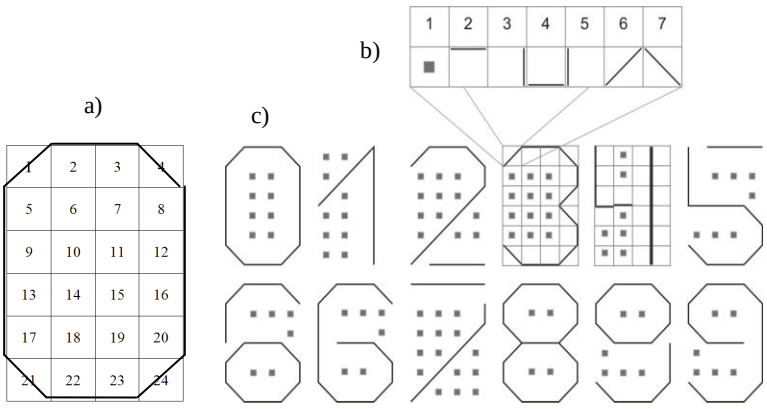

Рис. 3. Кодировка цифр

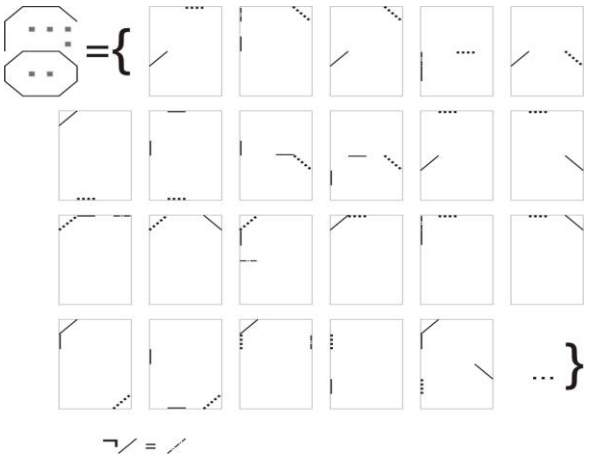

Рис. 4. Неподвижная точка цифры 6.

указания, где какая цифра). На этом множестве семантическим вероятностным выводом было обнаружено 55089 максимально специфических закономерностей – общих утверждений об объектах, о которых говорит Дж. Ст. Миллем. По этим закономерностям было обнаружено ровно 12 неподвижных точек, точно соответствующих цифрам.

Пример неподвижной точки для цифры 6 приведен на рис. 4. Рассмотрим, что представляет собой эта неподвижная точка. Занумеруем признаки цифр, как указано на рис. 3. Первая закономерность цифры 6 рис. 4, представленная в первом прямоугольнике после фигурной скобки означает, что, если в квадрате 13 стоит признак 6 (обозначим это как 13-6), то в квадрате 3 должен стоять признак 2 (обозначим как (3-2)). Предсказываемый признак обозначается точечной линией. Запишем эту закономерность как (13-6 $\Rightarrow$ 3-2). Нетрудно проверить, что эта закономерность выполнена на всех цифрах. Вторая закономерность означает, что из признака (9-5) и отрицания значения 5 первого признака $\neg$(1-5) (первый признак не должен быть равен 5) следует признак (4-7). Отрицание обозначается на рисунке пунктирной линией, как показано в нижней части рис. 4. Получим закономерность (9-5&$\neg$(1-5) $\Rightarrow$ 4-7). Последующие 3 закономерности в первой строке цифры 6 будут соответственно (13-6 $\Rightarrow$ 4-7), (17-5&$\neg$(13-5) $\Rightarrow$ 4-7), (13-6 $\Rightarrow$ 16-7).

На рис. 4 видно, что закономерности и признаки цифры 6 образуют неподвижную точку – взаимно предсказывают друг друга. Заметим, что закономерности используемые в неподвижной точке, выполнены на всех цифрах, а сама неподвижная точка выделяет только одну цифру. Это иллюстрирует феноменологическое свойство 2 ‘differences that make a difference’, в котором система причинных связей воспринимает «осознает» целостный объект. Поэтому цифры выделяются не закономерностями самими по себе, а их системной взаимосвязью.

Неподвижная точка формирует «прототип» по Eleanor Rosch или «образ» по Дж. Ст. Миллю. Программа не знает заранее, какие сочетания признаков максимально коррелируют между собой.

7. Понятие цели и теория функциональных систем

Рассмотрим целенаправленное поведение. Оно возникает, когда организм начинает двигаться. Когда он двигается, то он стремится изменить поступающую к нему стимуляцию, пищу и энергию нужным образом. Поэтому *любое действие с необходимостью становится целенаправленным* – оно стремится изменить приходящую к организму стимуляцию. Цель нельзя достичь, не имея критерия её достижения, иначе всегда можно считать, что цель уже достигнута и продолжать действие не нужно. Поэтому с необходимостью должен существовать критерий достижения цели, являющийся критерием остановки действия.

Понятие цели парадоксально – цель ничего не говорит о том, *как* её достичь и *как* надо организовать целенаправленное поведение. Для того, что бы знать, как достичь цель, нужен опыт. Если нет знаний и опыта, то поведение организуется методом «проб и ошибок». Для

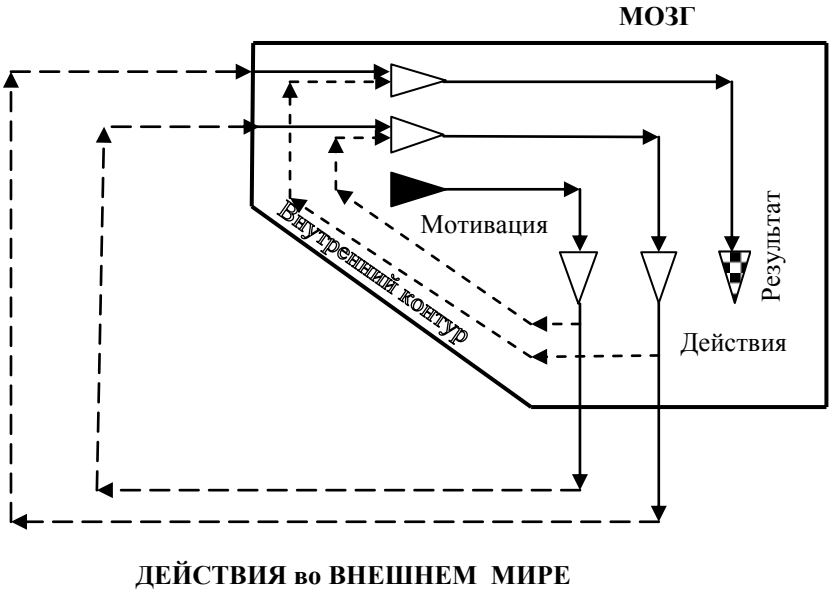

Рис. 5. Формирование акцептора результатов действия.

организации поведения методом «проб и ошибок» существует специальная ориентировочно-исследовательская реакция. Опыт складывается из полученных методом «проб и ошибок» случаев достижения цели. Достижение целей детально описывается в теории функциональных систем (Судаков 1984), формальная модель которой приведена в следующем разделе.

Причинность и нейрофизиологическое обеспечение акцептора результатов действий. В целенаправленном поведении достижение целей осуществляется на основе сделанных предсказаний. Для этого мозг должен уметь обнаруживать причинные связи.

Нейрофизиологически предвосхищение реализуется специальными коллатеральными ответвлениями от произведенных действий, которые поступают на «вход» мозга, конвергируя с афферентацией от входных стимулов (Судаков 1984: 97). Фактически это означает выработку условных (причинных) связей между осуществлением действий (эффекторным возбуждением) и последующим восприятием результатов действий, представленных их афферентными признаками (см. рис. 5). Осуществляя действия, мы сразу же по коллатералям посылаем условный сигнал о том, что сейчас получим афферентацию о результатах этих действий. Это приводит к выработке условных (причинных) связей между действиями и их результатами, отражающими связи действий и результатов во внешнем мире. Эти условные связи, осуществляемые мозгом по внутреннему контуру (см. рис. 5), позволяют прогнозировать результаты действий, происходящих во внешнем мире, ещё до появления самих результатов. Когда мотивационным возбуждением активируются различные последовательности действий, по достижению поставленной цели, то одновременно по «внутреннему контуру» прогнозируется вся последовательность и иерархия результатов, которые будут получены в процессе достижения цели. Когда принято решение об определенном плане действий, то одновременно по «внутреннему контуру» предвосхищается достижение всех промежуточных результатов, которые составляют акцептор результатов действия.

8. Формальная модель

Данная формальная модель продолжает работы (Анохин at al 2002, Витяев 2008, Демин Витяев 2008, Мухортов Хлебников Витяев 2012, Vityaev 2015) по формализации информационных процессов работы мозга, основанные на теории функциональных систем.

Приведем формальную модель, суммирующую упомянутые принципы и законы. Эта модель следующим образом учитывает приведенные рассмотрения:

1. использует формальную модель нейрона, обнаруживающую причинные связи и осно-

ванную на семантическом вероятностном выводе;
2. осуществляет постановку цели в целенаправленном поведении, формирует функциональную систему и акцептор результатов действия, который сверяют достигнутые результаты с ожидаемыми;
3. автоматически формирует подцели и подкрепляет достижение подцелей, если их достижение увеличивает вероятность достижения конечной цели;
4. моделирует сензорные коррекции Н.А. Бернштейна (Бернштейн 1997);
5. выбирает действие в реальном режиме времени с учетом текущей ситуации и получаемой афферентации;
6. планирует достижение цели в соответствии с последовательностью и иерархией функциональных систем по достижению всех результатов, требуемых для достижения конечной цели;
7. принимает решение об определенном способе достижения цели.

Будем предполагать, что эта модель является системой управления некоторого анимата, функционирующей в дискретном времени $t = 0, 1, ...,$. Пусть анимат имеет некоторый набор сенсоров $S_1, ..., S_n$, характеризующих состояние, как самого анимата, так и внешней среды. Каждый сенсор S_i имеет некоторое множество возможных показаний сенсора VS_i. Анимат также располагает множеством возможных действий в среде $A = \{a_1, ..., a_m\}$. Любое действие анимата, совершаемое в момент времени t_i, может приводить, в следующий момент времени $t_i + 1$ к какому-то изменению среды, и как следствие, к изменению показаний его сенсоров.

Поскольку анимат «воспринимает» окружающий мир только через свои сенсоры, то, с точки зрения анимата, состояние системы в каждый конкретный момент времени может быть записано в виде вектора показаний всех сенсоров $V(t) = (v_1, ..., v_n)$, где $v_i \in VS_i$ – показание *i*-го сенсора в момент времени *t*, причем состояния с одинаковыми показаниями сенсоров для анимата неразличимы. Множество всех возможных состояний системы обозначим как $S = (VS_1 \times VS_2 \times ... \times VS_n)$.

Поскольку, в общем случае, сенсоры анимата не могут учитывать всех физических законов среды и имеют собственные физические ограничения (например, по чувствительности, радиусу действия и т.п.), то при совершении аниматом некоторого действия в состоянии $s \in S$, система, с точки зрения анимата, может переходить в одно или несколько возможных состояний. Тогда, действие a_i анимата можно определить как функционал, переводящий систему «анимат – внешняя среда» из одного состояние в другое с некоторой вероятностью:

$a_i : (S_i) \to (S_i \times S \times P)$,

где S_i – подмножество S состояний системы, в которых действие a_i имеет смысл (осуществимо), $S_i \times S \times P$ – множество троек (s_0, s, p), где $s \in S$ – полученное в результате действия состояние, $p \in [0, 1]$ – вероятность его достижения из состояния $s_0 \in S_i$ при совершении действия a_i, вычисляемая в соответствии с объективными факторами осуществления действия во внешнем мире.

Определим понятие события и истории событий. Под событием $e = (s_0, s_e, a)$ будем понимать единичный факт перевода системы из состояния $s_0 \in S_0$ в состояние $s_e \in S$ в результате совершения действия a. Тогда историей H назовем множество пар (e, t), где e – событие, t – момент времени, когда произошло данное событие.

Теперь от общей модели «анимат-внешняя среда» перейдем к более конкретной дискретной модели. На множестве состояний системы $S = (VS_1 \cup VS_2 \cup ... \cup VS_n)$ определим множество предикатов $PS = \{P_1, ..., P_k\}$, каждый из которых вычисляется на основе показаний сенсоров. Каждое состояние системы, таким образом, может быть записано в виде вектора значе-

ний предикатов из PS, $s=(p_1,...,p_k)$, $p_i\in\{0,1\}$, где 1 означает истинность соответствующего предиката, а 0 – его ложность.

Задачей анимата является достижение некоторой цели. Определим цель *Goal* как состояние системы $s_{Goal}=(p_{i_1}^{goal},...,p_{i_{goal}}^{goal})$, которое требуется достичь. Запись $(p_{i_1}^{goal},...,p_{i_{goal}}^{goal})$ означает, что предикаты $p_{i_1}^{goal},...,p_{i_{goal}}^{goal}$ при достижении цели должны быть истинны.

Уточним понятие события и истории. Под событием $e=(s_0,s_e,a)$, как и раньше, будем понимать единичный факт перевода системы из состояния $s_0=(p_1^0,...,p_k^0)$ в состояние $s_e=(p_1^e,...,p_k^e)$ в результате совершения действия a, а под историей H событий – множество пар (e_t,t), где $e_t=(s_t,s_{t+1},a)$ – событие, t – момент времени, когда произошло данное событие.

Правила R, предсказывающие изменение состояния после осуществления действия *a* по внутреннему контуру работы мозга (рис. 5), определим как преобразование $R=(s_0\xrightarrow[p]{a}s_e)$, где:

s_0 – начальное состояние системы $(p_{i_1}^0,...,p_{i_0}^0)$;

s_e – конечное состояние системы $(p_{i_1}^e,...,p_{i_e}^e)$;

a – действие, которое переводит начальное состояние в конечное;

p – вероятность, с которой действие переводит начальное состояние в конечное.

Вероятность правила R рассчитывается следующим образом: если *n* – число случаев, когда начальным состоянием было s_0 и выполнялось действие *a*, а *m* – число тех случаев из *n*, когда действие a переводило состояние s_0 в состояние s_e, тогда $p=m/n$. Вероятности правил R (предсказывающие переход из состояния s_0 в состояние s_e после осуществлении действия *a* по внутреннему контуру мозга) и вероятности из множества P (предсказывающие переход из состояния s_0 в состояние s_e при осуществлении действия *a*, вычисляемые в соответствии с объективными факторами осуществления действия во внешнем мире) – различные величины. Можно сказать, что задачей обучения является максимальное приближение «субъективных» вероятностей правил R, оцениваемых аниматом, к объективным вероятностям P, характеризующим взаимодействие анимата с внешней средой.

Обнаружение правил осуществляется нейронами, замыкающие условные связи по внутреннему контуру работы мозга, в соответствии с семантическим вероятностным выводом.

Определим функциональную систему FSC, реализующую сенсорные коррекции, как набор $\mathrm{FSC}=(\mathrm{s_{Goal}},\mathrm{R_1},...,\mathrm{R_n},\mathrm{p_{FSC}})$. Функциональная система FSC осуществляет преобразование $s_0\xrightarrow[p_{FSC}]{R_1,...,R_n}s_{Goal}$, где $s_{Goal}=(p_{i_1}^{goal},...,p_{i_{goal}}^{goal})$ – целевое состояние функциональной системы, $R_1,...,R_n$ – правила вида $s_0\xrightarrow[p]{a}s_{Goal}$, с помощью которых из различных начальных состояний s_0 с помощью некоторого действия *a* можно попасть в целевое состояние s_{Goal} (рис. 6). Цель s_{Goal} функциональной системы ставится соответствующим мотивационным возбуждением. Способ вычисления вероятности p_{FSC} приведен ниже.

В соответствии с принципом сензорных коррекций Н.А.Бернштейна, принципиально нельзя знать заранее точный результат предыдущего движения. Поэтому выбрать максимально вероятное правило $s_0\xrightarrow[p]{a}s_{Goal}$, приводящее к достижению цели, в текущем состоянии $s_t=(p_1^t,...,p_k^t)$ можно только после поступления афферентация о завершении предыдущего действия, чтобы выбрать правило с начальным состоянием $s_0=(p_{i_1}^0,...,p_{i_0}^0)$, соответствующим текущему состоянию $\{p_{i_1}^0,...,p_{i_0}^0\}\subset\{p_1^t,...,p_k^t\}$ (обозначим это как $s_0\subseteq s_t$, на рис. 6).

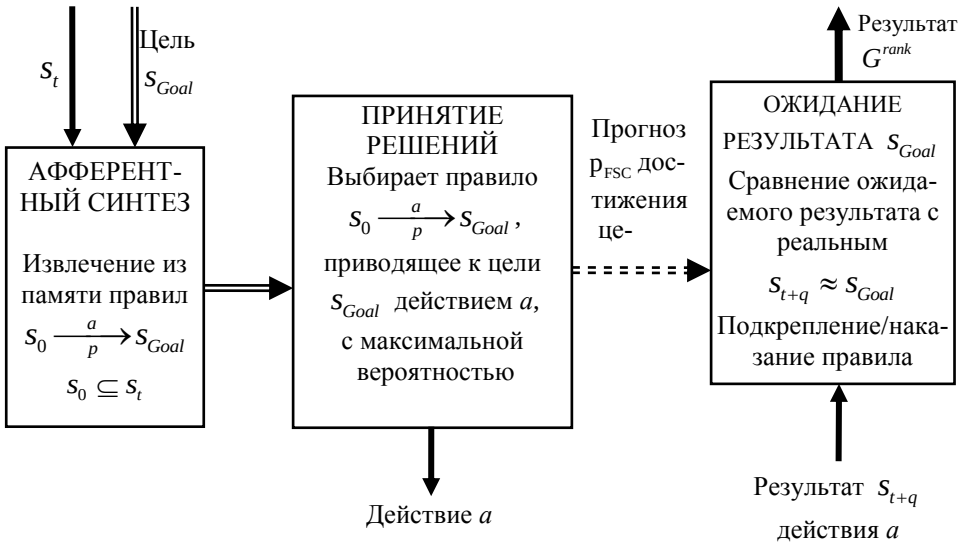

Рис. 6. Схема функциональной системы, реализующей сенсорные коррекции.

Когда функциональной системой верхнего уровня, удовлетворяющей некоторую потребность, принимается решение и перебираются различные последовательности/иерархии действий по достижению цели, то мы также принципиально не можем знать тех состояний s_t, которые возникнут в результате реального осуществления этой последовательности/иерархии действий. Мы также не можем знать, какие будут выбраны правила для достижения цели каждого конкретного действия в этой последовательности/иерархии. Тем не менее, для принятия решения необходим прогноз вероятности достижения цели. Оценку вероятности достижения цели функциональной системой FSC можно подсчитать, опираясь на статистику достижения цели следующим образом: если n – число случаев, когда поступил запрос на достижение цели s_{Goal}, а m – число случаев, когда выбранные правила и последовательности/иерархии действий привели к достижению цели s_{Goal}, то $p_{FSC} = m/n$. Поэтому на рис. 6 прогноз достижения цели осуществляется с вероятностью p_{FSC} достижения цели функциональной системой.

Когда в момент времени *t* пришел запрос на достижение цели s_{Goal} функциональной системой $FS\tilde{N}$ в текущем состоянии $s_t = (p_1^t, ..., p_k^t)$, то она:

1. выбирает правило $s_0 \xrightarrow[p]{a} s_{Goal}$ из набора $R_1, ..., R_n$, которое:
   a. применимо в текущей ситуации $s_0 \subseteq s_t$;
   b. может достичь цели s_{Goal} с максимальной вероятностью *p*;
2. ожидает в акцепторе результатов действия достижение цели s_{Goal} после осуществления действия *a*;
3. сравнивает акцептором результатов действия достигнутое состояние $s_{t+q} = (p_1^{t+q}, ..., p_k^{t+q})$ в момент $t+q$, в результате осуществления действия *a*, с целью $s_{Goal} \approx s_{t+q}$. Если $s_{Goal} \subset s_{t+q}$, то цель достигнута и правило $s_0 \xrightarrow[p]{a} s_{Goal}$ подкрепляется (его статистика увеличивается);
4. если для текущего состояния s_t нет подходящего правила, либо после применения выбранного правила и соответствующего ему действия целевое состояние s_{Goal} не достигнуто, то функциональной системой FSC в ответ на запрос возвращается, что цель не достигнута и выбранное правило наказывается (его статистика уменьшается).

Функциональные системы в общем случае являются последовательностями и иерархией функциональных систем FSC, реализующих рефлекторные кольца.

Функциональной системой FS, объединяющей последовательность функциональных систем вида FSC, будет набор $\mathrm{FS} = (\mathrm{s_{Goal}}, \mathrm{FSC_1}, ..., \mathrm{FSC_n}, \mathrm{p_{FS}})$, реализующий преобразование $FS = s_0 \xrightarrow[\rightarrow s_1 \rightarrow s_2 \rightarrow ... \rightarrow s_{goal} \quad p_{FS} = p_{FSC_1} \bullet ... \bullet p_{FSC_n}]{FSC_1, ..., FSC_n} s_{goal}$, где $FSC_1 = (s_0 \xrightarrow[p_{FSC_1}]{R_1^1, ..., R_{v_1}^1} s_1)$,

$FSC_2 = (s_0 \xrightarrow[p_{FSC_2}]{R_1^2, ..., R_{v_2}^2} s_2), \ldots,$ $FSC_n = (s_0 \xrightarrow[p_{FSC_n}]{R_1^n, ..., R_{v_n}^n} s_{goal})$ – функциональные системы рефлекторных колец. Цель функциональной системы FS состоит в последовательном достижении целей $s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow ... \rightarrow s_{goal}$ функциональными системами $FSC_1, ..., FSC_n$ с суммарной вероятностью $p_{FS} = p_{FS\tilde{N}_1} \bullet ... \bullet p_{FS\tilde{N}_n}$. Такие функциональные системы могут образовываться автоматически, как это описано ниже.

Функциональные системы FS, объединяющие последовательности и иерархии функциональных систем FSC, возникают в случае, когда в последовательностях функциональных систем FSC встречаются также функциональные системы FS. Тогда функциональная система $\mathrm{FS} = (\mathrm{s_{Goal}}, \mathrm{FS_1^1}, ..., \mathrm{FS_n^1}, \mathrm{p_{FS}})$ есть последовательность функциональных систем, реализующих преобразование

$$FS = s_0 \xrightarrow[\rightarrow s_1 \rightarrow s_2 \rightarrow ... \rightarrow s_{goal} \quad p_{FS} = p_{FS_1^1} \bullet ... \bullet p_{FS_n^1}]{FS_1^1, ..., FS_n^1} s_{goal},$$

где $\mathrm{FS_i^1}$ – либо FS, либо $FS\tilde{N}$. Например, если $\mathrm{FS_i^1}, \mathrm{FS_j^1} \in \{\mathrm{FS_1^1}, ..., \mathrm{FS_n^1}\}, \mathrm{i < j}$ реализуют преобразования

$$FS_i^1 = \xrightarrow[\rightarrow s_1^i \rightarrow s_2^i \rightarrow ... \rightarrow s_i \quad p_{FS_i^1}]{FS(i)_1^2, ..., FS(i)_{n_i}^2} s_i, \quad FS_j^1 = \xrightarrow[\rightarrow s_1^j \rightarrow s_2^j \rightarrow ... \rightarrow s_j \quad p_{FS_j^1}]{FS(j)_1^2, ..., FS(j)_{n_j}^2} s_j,$$

то функциональные системы $\mathrm{FS(i)_1^2}, ..., \mathrm{FS(i)_{n_i}^2}, \mathrm{FS(j)_1^2}, ..., \mathrm{FS(j)_{n_j}^2}$ находятся уже на уровне 2 и преобразование, реализуемое функциональной системой FS, имеет вид

$$FS = s_0 \xrightarrow[\rightarrow s_1 \rightarrow s_2 \rightarrow ... \rightarrow [\rightarrow s_1^i \rightarrow s_2^i \rightarrow ... \rightarrow s_i] ... \rightarrow [\rightarrow s_1^j \rightarrow s_2^j \rightarrow ... \rightarrow s_j] ... \rightarrow s_{goal} \quad p_{FS}]{FS_1^1, ... FS_i^1[FS(i)_1^2, ..., FS(i)_{n_i}^2], ..., FS_j^1[FS(j)_1^2, ..., FS(j)_{n_j}^2], ..., FS_n^1} s_{goal}.$$

Каждая функциональная система представляет собой тот или иной способ достижения цели s_{Goal}, В соответствии с теорией организации движений Н.А.Бернштейна, ведущим уровнем организации движений является верхний уровень $\mathrm{FS} = (\mathrm{s_{Goal}}, \mathrm{FS_1^1}, ..., \mathrm{FS_n^1}, \mathrm{p_{FS}})$ ранга 1, соответствующий смыслу решаемой задачи. Функциональная система верхнего уровня может вызывать функциональные системы более низких уровней.

Когда приходит запрос на достижение цели s_{Goal} функциональной системой FS, то она:

1) выбирает правила, применимые в текущей ситуации, для первой из функциональных систем FSC, входящих в данную функциональную систему. Если для текущего начального состояния s_0 первой FSC нет подходящего правила, то функциональная система FS не применима к данной ситуации;
2) формирует «конкретную цель» (высшую мотивацию) в виде последовательности и иерархии целей всех, входящих в нее, функциональных подсистем. Например, для приведенной выше функциональной системы это будет последовательность
$$s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow ... \rightarrow [\rightarrow s_1^i \rightarrow s_2^i \rightarrow ... \rightarrow s_i] ... \rightarrow [\rightarrow s_1^j \rightarrow s_2^j \rightarrow ... \rightarrow s_j] ... \rightarrow s_{goal}.$$
3) прогнозирует достижение цели s_{Goal} с вероятностью $\mathrm{p_{FS}}$;
4) ожидает (акцептором результатов действия) достижение всей последовательности и иерархии целей всех входящих в нее FSC после выполнения соответствующих действий;
5) запускает последовательное выполнение действий в функциональных подсистемах $FS\tilde{N}$;
6) если в какой-либо функциональной подсистеме цель не достигнута, что фиксируется акцептором результатов действий этой функциональной системы, то возникает ориен-

тировочно-исследовательская реакция, которая выбирает другую функциональную систему FS для достижения цели s_{Goal}. Правила этой функциональной подсистемы наказываются;

7) достижение результата каждой функциональной подсистемой фиксируется акцептором результатов действия и подкрепляется.

Опишем все элементы архитектуры функциональных систем, используя введенные определения.

Афферентный синтез включает в себя синтез мотивационного возбуждения, памяти, обстановочной и пусковой афферентации, а также обратную афферентацию об осуществленных действиях, приходящую по коллатералям пирамидного тракта. Вся эта афферентация может быть задана набором сенсоров $S_1, ..., S_n$, включая сенсоры *мотивационного возбуждения, обстановочной и пусковой афферентаций*. Мотивационным возбуждением также задается цель $Goal = (p_{i_1}^{goal}, ..., p_{i_{goal}}^{goal})$.

Память. Каждая цель может достигаться различными последовательностями действий, реализуемыми различными функциональными системами. Поэтому мотивация извлекает из памяти все функциональные системы $\mathrm{FS} = (\mathrm{s_{Goal}}, \mathrm{FS_1^1}, ..., \mathrm{FS_n^1}, \mathrm{p_{FS}})$, приводящие к достижению этой цели.

Обстановочная и пусковая афферентации задают текущее состояние системы $s_t = (p_1, ..., p_k)$ в каждый момент времени *t*. Начальные состояния $s_0 = (p_{i_1}^0, ..., p_{i_0}^0)$ применяемых в этот момент правил $s_0 \xrightarrow[p]{a} s_e$ должны соответствовать текущему состоянию системы $s_0 \subseteq s_t$.

«Вытягивая» из памяти весь накопленный опыт, мотивационное возбуждение как цель преобразуется в ***конкретную цель*** «высшую мотивацию», определяющую способ своего достижения. Для каждой функциональной системы $\mathrm{FS} = (\mathrm{s_{Goal}}, \mathrm{FS_1^1}, ..., \mathrm{FS_n^1}, \mathrm{p_{FS}})$ конкретной целью является вся последовательность и иерархия целей всех входящих в нее функциональных подсистем, например

$$s_0 \to s_1 \to s_2 \to ... \to [\to s_1^i \to s_2^i \to ... \to s_i] ... \to [\to s_1^j \to s_2^j \to ... \to s_j] ... \to s_{goal}.$$

Принятие решения. На стадии афферентного синтеза мотивационным возбуждением может быть извлечено из памяти множество функциональных систем $\mathrm{FS} = (\mathrm{s_{Goal}}, \mathrm{FS_1^1}, ..., \mathrm{FS_n^1}, \mathrm{p_{FS}})$, достигающих цель $\mathrm{s_{Goal}}$. На стадии принятия решения выбирается

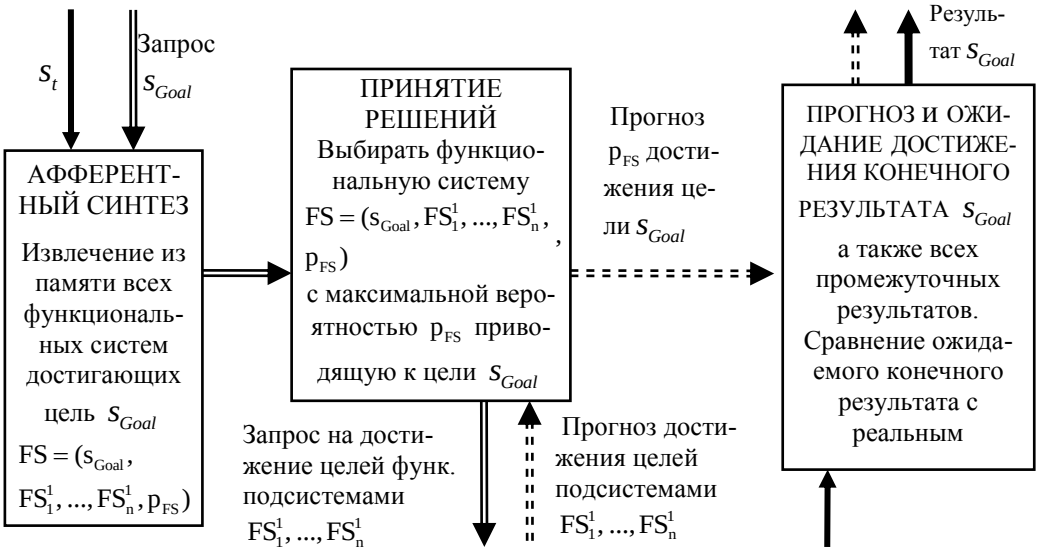

Рис. 7. Схема функциональной системы.

одна из них и фиксируется *конкретный план действий*. Процесс принятия решений осуществляется переключающей функции эмоций (см. рис. 7).

Акцептор результатов действия. Мотивационное возбуждение, преобразуясь в конкретную цель, извлекает из памяти также и конкретный критерий достижения цели – *акцептор результатов действия*, который состоит из всей совокупности критериев по достижению всей последовательности и иерархии целей

$$s_0 \to s_1 \to s_2 \to ... \to [\to s_1^i \to s_2^i \to ... \to s_i] ... \to [\to s_1^j \to s_2^j \to ... \to s_j] ... \to s_{goal} .$$

Заключение

Суммируя вышесказанное, можно выделить следующие основные функции сознания:

1. Обеспечение логически непротиворечивого и прогностического представления реальности. Это становится возможным за счет хорошей структурированности самого внешнего мира – его причинности, «естественной» классификации и «встроенности». Мозг улавливает эту структурированность клеточными ансамблями и внутренним контуром работы мозга в процессе деятельности, которые в нашей формализации представлены формальной работой нейрона, обнаруживающей причинные связи, а также неподвижными точками предсказаний, образующими логически непротиворечивые модели реальности, включающие как стимулы, которые возможны в данной модели, так и стимулы, которые в данной модели не возможны. Это создает конкуренцию между возможными моделями реальности, включая предметность воспринимаемого образа (Столин 1976), заставляя сознание искать наиболее непротиворечивую модель («образ мира») реальности («видимом поле»).
2. В создаваемый в данный момент «образ мира», сознание включает, прежде всего, потребности и вероятности их удовлетворения, что вызывает соответствующие эмоции и субъективные состояния (qualia). В вероятностный прогноз не входят автоматизмы действий, поэтому они уходят из сознания.
3. Другой (по отношению к непротиворечивости) основной функцией сознания является постоянная и непрерывная во времени и пространстве проверка совпадения, создаваемой модели реальности («образа мира») и самой реальности. Это осуществляется постоянной проверкой во времени и пространстве совпадения прогнозов, осуществляемых моделью реальности с самой реальностью, что создает ощущение внешнего мира. Поэтому ощущения концентрируются в местах прогноза. В нашей формализации и модель нейрона и неподвижные точки, а также неподвижные точки, включающие деятельность, осуществляют прогноз. Неподвижная точка замыкает прогнозы внутри себя только в том случае, когда все они подтвердились, если нет, то в неподвижной точке возникает противоречие, которое вызывает ориентировочно-исследовательскую реакцию и изменение стимуляции.

Литература

Анохин К.В., Бурцев М.С., Зарайская И.Ю., Лукашев А.О., Редько В.Г. Проект «Мозг анимата»: разработка модели адаптивного поведения на основе теории функциональных систем // Восьмая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. М.: Физматлит, 2002. Т.2. С.781-789.

Бернштейн Н.А. Биомеханика и физиология движений. // Избранные психологические труды, Москва-Воронеж, 1997, с.605.

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, Институтом математики им. С. Л. Соболева СО РАН, 2015, стр. 1006-1031.

Витяев Е.Е., Принципы работы мозга, содержащиеся в теории функциональных систем П.К. Анохина и теории эмоций П.В. Симонова // Нейроинформатика, 2008, том 3, № 1, стр. 25-78

Витяев Е.Е., Неупокоев Н.В. Математическая модель восприятия и образа. Информационные технологии в гуманитарных исследованиях, Вып.17, ИАЭТ СО РАН, Новосибирск, 2012, 63-72.

Витяев Е.Е., Неупокоев Н.В. Формальная модель восприятия и образа как неподвижной точки предвосхищений // Подходы к моделированию мышления. УРСС Эдиториал, Москва, 2014, стр. 155-172.

Витяев Е.Е., Демин А.В., Пономарёв Д.К. Вероятностное обобщение формальных понятий // Программирование, Т.38, №5, 2012, С. 219-230.

Гибсон Дж. Экологический подход к зрительному восприятию. М.: Прогресс, 1988. С. 462.

Демин А.В., Витяев Е.Е. Логическая модель адаптивной системы управления. Нейроинформатика, 2008, том 3, № 1, стр. 79-107

Забродин В.Ю. О критериях естественной классификации // НТИ, сер.2, 1981, №8.

Закон. Необходимость. Вероятность. М., «Прогресс», 1967, с.366

Рудольф Карнап. Философские основания физики. М., «Прогресс», 1971, с.388

Леонтьев А.Н. Образ мира // Избранные психологические произведения. – М.: Педагогика, 1983. – С. 251-261.

Мейен С.В., Шрейдер С.А. Методологические аспекты теории классификаций // Вопросы философии, 1976, №12

Мухортов В.В., Хлебников С.В., Витяев Е.Е. Улучшенный алгоритм семантического вероятностного вывода в задаче 2-мерного анимата // Нейроинформатика, 2012, том 6, № 1, стр. 50-62

Найсер У. Познание и реальность. “Прогресс”, М. 1981, с. 229.

В.Г.Редько Эволюция, нейронные сети, интеллект. Модели и концепции эволюционной кибернетики. М., «ЛИБРОКОМ», 2011, с.220

Симонов П.В. Эмоциональный мозг. М.: Наука, 1981. с. 140.

Симонов П.В. Высшая нервная деятельность человека (мотивационно-эмоциональные аспекты). М.: Наука, 1975. с. 173.

Смирнов Е.С. Конструкция вида таксономической точки зрения // Зоол. Журн. Т. 17, №3, 1938, С. 387-418.

Смирнов С.Д. Психология образа: проблема активности психического отражения. МГУ, М., 1985, с.232.

Столин В.В. Исследование порождения зрительного пространственного образа. — В кн.: Восприятие и деятельность. М., 1976.

Судаков К.В. Общая Теория Функциональных Систем М.: Медицина, 1984. с. 222.

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. Cognition, 69, 135–178.

Hebb D.O. The organization of behavior. A neurophysiological theory. NY, 1949. 335 p.

Hempel, C. G. ‘Maximal Specificity and Lawlikeness in Probabilistic Explanation’, Philosophy of Science 35, 1968. – P. 16–33.

Masafumi Oizumi, Larissa Albantakis, Giulio Tononi. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology, May 2014, V.10. Issue 5.

Mill, J.S. System of Logic, Ratiocinative and Inductive. L., 1843.

Bob Rehder. Categorization as causal reasoning // Cognitive Science 27 (2003) 709–748.

Bob Rehder, Jay B. Martin. 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 pp. 2944-2949.

Rosch, E., Mervis, C.B. Family resemblances. Studies in the internal structure of categories // Cognitive Psychology, 7, 1975, P. 573–605.

Rosch, E., Principles of Categorization // Rosch, E. & Lloyd, B.B. (eds), Cognition and Categorization, Lawrence Erlbaum Associates, Publishers, (Hillsdale), 1978. P. 27–48

B. H. Ross, E. G. Taylor, E. L. Middleton, and T. J. Nokes. Concept and Category Learning in Humans // H. L. Roediger, III (Ed.), Cognitive Psychology of Memory. Vol. [2] of Learning and Memory: A Comprehensive Reference, 4 vols. (J.Byrne Editor), Oxford: Elsevier, 2008, P. 535-556.

The Nature of Classification. Relationships and Kinds in the Natural Sciences. Palgrave Macmillan. 2013. 208p.

Evgenii Vityaev. The logic of prediction // Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference (August 16-19, 2005, Novosibirsk, Russia), edited by S.S. Goncharov, R. Downey, H. Ono, World Scientific, Singapore, 2006, P. 263-276.

Vityaev E.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 (A. Chella, R.Pirrone, R. Sorbello, K.R. Johannsdottir, Eds). In Advances in Intelligent Systems and Computing, v.196, Springer: Heidelberg, New York, Dordrecht, London. 2013, P. 339-344.

Evgenii Vityaev. Unified formalization of "natural" classification, "natural" concepts, and consciousness as integrated information by Giulio Tononi // The Sixth international conference on Biologically Inspired Cognitive Architectures (BICA 2015, November 6-8, Lyon, France), Procedia Computer Science, v.71, Elsevier, 2015. pp 169-177.

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.

E.E. Vityaev, A.V. Demin, D. K. Ponomaryov. Probabilistic Generalization of Formal Concepts // Programming and Computer Software, 2012, Vol. 38, No. 5. P. 219–230.

E.E. Vityaev, L.I. Perlovsky, B.Y. Kovalerchuk, S.O. Speransky Probabilistic dynamic logic of cognition. Biologically Inspired Cognitive Architectures. Special issue: Papers from the Fourth Annual Meeting of the BICA Society (BICA 2013), v.6, October, Elsevier, 2013, pp.159-168.

E.E. Vityaev, V.V. Martinovich. Probabilistic Formal Concepts with Negation // A. Voronkov, I. Virbitskaite (Eds.): PCI 2014, LNCS 8974, P. 385-399.

Витяев Евгений Евгеньевич
Ведущий научный сотрудник, институт математики им. С.Л.Соболева, Новосибирск
vityaev@math.nsc.ru