УДК 577.38*004.81
DOI 10.17726/phillT.2023.2.9

Биофизический подход к моделированию рефлексии: обоснование, методы, результаты[1]

Барцев Сергей Игоревич,

доктор физико-математических наук,
профессор, кафедра биофизики,
Институт фундаментальной биологии и биотехнологий,
ФГАОУ ВО «Сибирский федеральный университет»
Красноярск, Россия;

главный научный сотрудник,
лаборатория теоретической биофизики,
Институт биофизики Сибирского отделения Российской академии наук – обособленное подразделение федерального государственного бюджетного научного учреждения Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения Российской академии наук»
Красноярск, Россия

bartsev@yandex.ru

Маркова Галия Муратовна,

аспирант, ассистент, кафедра биофизики,
Институт фундаментальной биологии и биотехнологий,
ФГАОУ ВО «Сибирский федеральный университет»
Красноярск, Россия;

лаборант, лаборатория теоретической биофизики,
Институт биофизики Сибирского отделения Российской академии наук – обособленное подразделение федерального государственного бюджетного научного учреждения Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения Российской академии наук»
Красноярск, Россия

GMarkova@ibp.ru

[1] Работа поддержана грантом РНФ № 23-21-10041, Красноярского краевого фонда науки «Иерархия функциональных аттракторов в нейросетевых моделях рефлексии».

Матвеева Алевтина Игоревна,
аспирант,
Институт биофизики Сибирского отделения Российской академии наук – обособленное подразделение федерального государственного бюджетного научного учреждения Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения Российской академии наук»
Красноярск, Россия

matveevaalevtinai@gmail.com

Аннотация. Используемый физикой подход, основанный на выделении и исследовании идеальных объектов, лежащий также в основе биофизики в сочетании с эвристическим моделированием фон Неймана и функциональным фракционированием по Р. Розену, обсуждается в качестве инструмента исследования свойств сознания. Объектом исследования становится своеобразная линейка систем-аналогов: человеческий мозг, мозг позвоночных, мозг беспозвоночных и искусственные нейросети, способные осуществлять рефлексию, которая является ключевым свойством или характеристикой сознания. Рефлексия в широком смысле слова, понимаемая как внутреннее отображение внешнего мира, свойственна широкому кругу животных, причем некоторые из них (шмели, рыбы) демонстрируют даже рефлексию в узком смысле этого слова, понимаемую как внутреннее представление себя. Реализуется это сложное поведение с помощью миниатюрных мозгов ~1 млн нейронов. Проиллюстрировано использование простых рекуррентных нейронных сетей (РНС) для получения ответов на вопросы общего характера. Например, показано, что небольшая РНС способна проходить тест отложенного сравнения с образцом, формируя индивидуальную динамическую репрезентацию поступившего стимула, допускающую декодирование специальным нейронным детектором. Продемонстрировано, что в рефлексивной игре «чет-нечет» РНС имеет огромное преимущество над слоистой нейросетью, с тем же и большим количеством нейронов – рефлексия побеждает регрессию. Обнаружено, что асимметрия исходов в игре «чет-нечет», для объяснения которой привлекали различные причины, включая психологические («догонять легче, чем убегать»), воспроизводится в игре двух РНС. Очевидно, что психологические причины здесь отсутствуют и преимущество игрока, играющего за «чет», объясняется более сложной стратегией «нечет»-игрока: ему нужно предсказать ход противника и выбрать противоположный.

Ключевые слова: идеальные объекты; эвристическое моделирование; системы-аналоги; рекуррентность рефлексии; отложенный тест сравнения с образцом; рефлексия и регрессия; асимметрия игры «чет-нечет».

Biophysical approach to modeling reflection: basis, methods, results

Bartsev Sergey Igorevich,

Doctor of Physical and Mathematical Sciences, Professor,
Department of Biophysics,
School of Fundamental Biology and Biotechnology,
Siberian Federal University
Krasnoyarsk, Russia;

Chief Researcher,
Laboratory of Theoretical Biophysics,
Biophysics Institute of the Siberian Branch of the RAS –
Division of Federal Research Center
«Krasnoyarsk Scientific Center of the Siberian Branch of the RAS»
Krasnoyarsk, Russia

bartsev@yandex.ru

Markova Galiya Muratovna,

Postgraduate Student, Assistant,
Department of Biophysics,
School of Fundamental Biology and Biotechnology,
Siberian Federal University
Krasnoyarsk, Russia;
Laboratory Assistant,

Laboratory of Theoretical Biophysics,
Biophysics Institute of the Siberian Branch of the RAS –
Division of Federal Research Center
«Krasnoyarsk Scientific Center of the Siberian Branch of the RAS»
Krasnoyarsk, Russia

GMarkova@ibp.ru

Matveeva Alevtina Igorevna,

Postgraduate student,
Biophysics Institute of the Siberian Branch of the RAS –
Division of Federal Research Center
«Krasnoyarsk Scientific Center of the Siberian Branch of the RAS»
Krasnoyarsk, Russia

matveevaalevtinai@gmail.com

Abstract. The approach used by physics is based on the identification and study of ideal objects, which is also the basis of biophysics, in combination with von Neumann heuristic modeling and functional fractionation according to R.Rosen is discussed as a tool for studying the properties of consciousness. The object of the study is a kind of line of analog systems: the human brain, the vertebrate brain, the invertebrate brain and artificial neural networks capable of reflection, which is a key property characteristic of consciousness. Reflection in the broad sense of the word, understood as an internal representation of the external world, is characteristic of a wide range of animals, and some of them (bumblebees, fish) even demonstrate reflection in the narrow sense of the word, understood as an inner self-representation. This complex behavior is realized by miniature brains of ~1 million neurons. The use of simple recurrent neural networks (RNNs) to obtain answers to general questions is illustrated. For example, it has been shown a small RNS is able to pass delayed matching to sample (DMTS) test, forming an individual dynamic representation of the received stimulus, allowing decoding by a special external neural detector. . It has been demonstrated in the reflexive game "even-odd", the RNS has a huge advantage over a multi-layered neural network, with the same and a larger number of neurons – reflection defeats regression. It was found that the asymmetry of outcomes in the odd-even game, which was explained by various causes, including psychological ones – "it's easier to catch up than to run away", is reproduced in the game of two RNNs. Obviously, there are no psychological causes here and the advantage of the player playing for "even" is explained by the more complex strategy of the "odd" player – he needs to predict the opponent's move and choose the opposite one.

Keywords: ideal objects; heuristic modeling; analog systems; recurrence of reflection; delayed matching to sample test; reflection and regression; asymmetry of the even-odd game.

Введение

В настоящее время существует много (более 20) теорий сознания, базирующихся на различных философских и методологических основаниях [1-5], что означает наличие некоторого тупика в понимании природы сознания. В этой ситуации представляется естественным пытаться применять различные подходы и методы для продвижения в этой проблеме.

Возможности естественных наук в понимании природы сознания представляются достаточно ограниченными по двум причинам:

1) естественные науки работают с объективным миром, дающим возможность получать факты – инварианты эмпирического опыта, т.е. инварианты в отношении смены наблюдателя. В то же время сознание по природе своей субъективно и недоступно внешнему наблюдателю;

2) наука не занимается объяснением природы явлений, она может лишь строить математические описания – модели. По мнению А. Ю. Хренникова, «объяснить» взаимоотношения духа и материи также невозможно, как и, например, «объяснить» взаимоотношения вещества и электричества [6, с. 27]. В отличие от ситуации с сознанием Максвелл создал математическую модель, описывающую это взаимодействие. Тогда с позиции естественных наук проблема «сознание-материя» – это проблема построения математической модели, которая будет описывать ментально-физические процессы.

Современная когнитивная наука с подачи Ф. Крика и К. Коха [7-9] в основном отказалась от попыток дать научное определение сознания и занялась выявлением нейронных коррелятов сознания (НКС-концепция). Сложность в том, что попытки сопоставления различных феноменов сознания с нейрофизиологическими данными проводятся на сложнейшем материальном носителе – человеческом мозге, математическую модель которого вряд ли можно построить.

Выходом из создавшейся ситуации может быть параллельное движение сразу в двух направлениях:

1) перенос фокуса внимания на более простые системы, обладающие сознанием, и

2) изучение не всего многообразия феноменов сознания, а концентрация на некоторой ключевой характеристике сознания и изучение ее проявлений в условиях, гарантирующих либо ее наличие, либо отсутствие, что позволит выявить структуры, в которых она может проявиться, и найти условия ее реализации.

Оба эти направления характерны для биофизического подхода к изучению живых систем. И поэтому целью данной работы является обоснование применения биофизического подхода к феномену сознания, обсуждение его особенностей и связей с другими подходами к изучению сложных систем и демонстрация применения этого подхода к решению некоторых задач когнитивистики.

Методы и материалы

1. Что такое биофизика?

В первую очередь нужно описать специфику биофизического подхода. Не отвлекаясь на обсуждение различных определений биофизики [10, с. 4], сформулируем рабочее определение этой науки. За основу возьмем парадоксальное по форме определение, предложенное Л. А. Блюменфельдом: «Биофизика – это область биологии, в которой должны предпочтительно работать ученые, имеющие фундаментальное физическое образование» [11, с. 5]. По-видимому, он считал, что фундаментальное физическое образование имеет свои особенности и способ мышления человека, получившего это образование, отличается от способа мышления других специалистов.

Суть физического подхода исчерпывающе выразил известный биофизик Н. Рашевский: «Мы начинаем с исследования в высшей степени ***идеализированных систем***, которые могут не иметь никаких прямых аналогов в реальной природе. …Против такого подхода можно выдвинуть возражение, что подобные системы не имеют никакой связи с действительностью и что поэтому никакие заключения относительно таких систем не могут быть перенесены на реальные системы. Тем не менее именно этот подход применяли и всегда применяют в физике. Физик занимается детальным математическим исследованием таких нереальных вещей, как «материальные точки», «абсолютно твердые тела», «идеальные жидкости» и т. п. ***В природе подобных вещей не существует.*** Однако же физик не только изучает их, но и применяет свои выводы к ***реальным вещам.*** И что же? Такое применение ведет к практическим результатам – по крайней мере, в известных пределах. Все дело в том, что в этих пределах реальные вещи имеют свойства, общие с воображаемыми идеальными объектами!» (цитируется с небольшими сокращениями по Моровиц [12, с. 41].

Для иллюстрации этого подхода приведем пример из школьного курса физики – математический маятник: материальная точка (объект, имеющий массу, но не имеющий размера, – **не бывает!**), подвешена на невесомой (**не бывает!**), нерастяжимой (**не бывает!**), бесконечно тонкой (**не бывает!**) нити, с нулевым сопротивлением изгибу (**не бывает!**). При малом угле отклонения α от положения равновесия ($\sin\alpha \approx \alpha$) можно получить простое уравнение свободных колебаний этого маятника. Но, что самое интересное,

полученное выражение для зависимости периода колебаний от длины подвеса и ускорения свободного падения: $T = 2\pi\sqrt{\frac{l}{g}}$, применимо к реальным физическим маятникам (при малых углах отклонения) и позволяет рассчитывать их параметры.

На основе вышесказанного можно сформулировать рабочее определение биофизики: ***биофизика – это наука, занимающаяся построением и исследованием идеализированных систем, моделирующих ключевые свойства живого на разных уровнях его организации.***

2. Эвристические модели и функциональное фракционирование

Если изучаемая система очень сложна, то выделить идеализированную подсистему очень непросто. В этом случае исследователи стараются найти природный модельный объект, упрощающий изучение требуемого общего свойства, а также могут обратиться к математическим моделям особого вида. Полезность перехода к искусственным модельным объектам обосновал фон Нейман: «Поскольку у нас нет достаточно ясного представления о том, как функционируют живые организмы, то обращение к органике большой пользы нам не принесет. Мы займемся поэтому автоматами, которые мы в совершенстве знаем, ибо мы их сделали. Опишем автоматы, способные воспроизводить себя» [13, с. 98].

Подход, основанный на построении абстрактных моделей, был назван Дж. фон Нейманом ***эвристическим методом***, сущность которого заключалась в том, что поиск решения на компьютере не является самоцелью, а ведется для того, чтобы выявить удобные понятия, широко приложимые принципы и построить общую теорию.

Для выбора адекватной эвристической модели необходимо решить, какое свойство или признак исследуемой системы является важным и в то же время общим для широкого класса биологических систем данного типа. По мнению Дж. Бернала, «биология методологически отличается от других естественных наук тем, что в фокусе внимания находятся, прежде всего, ***функционирование*** и эволюция систем. Структура здесь имеет значение только в связи с функцией и происхождением…» [14, с. 112].

Выделение Берналом функционирования, как особой харак-

теристики живого, согласуется с подходом Н. Рашевского и его ученика Р. Розена. Рашевский писал: «…Данному набору входов в некоторый орган соответствует определенный набор выходов. Если мы изменим какие-то входы, мы тем самым изменим выходы. Соответствие в математике называют *отображением.* Следовательно, мы можем сказать, что орган отображает множество входов на соответствующее множество выходов. Весь организм, таким образом, становится набором или системой отображений» [15, с. 63].

Вклад самого Розена в теорию сложных систем уникален в том смысле, что он подходил к биологической организации по сути нередукционистским способом. Вместо обсуждения физических объектов (генов, ферментов, органелл и т. д.) он рассматривал системные функции (метаболизм, самовоспроизводство, организационный инвариант) [16].

Розеном выдвинут тезис, что любое функциональное свойство данной системы может быть исследовано одинаково хорошо на любом из системных аналогов или даже целиком абстрактно. Такое абстрактное функциональное свойство, проявляемое каждым из системных аналогов, которые *реализуют* абстрактную систему, он назвал ***динамической метафорой.***

На этом пути развивается то, что может быть названо ***функциональным фракционированием,*** представляющим сложную систему в виде совокупности динамических метафор. Такие динамические метафоры могут играть важную роль в нашем понимании биологических процессов, в отличие от обыкновенного структурного моделирования [17].

Принимая, как это делал Р. Розен, функционирование за основу рассмотрения, мы делаем акцент на целостном представлении биологических систем, поскольку функция – это то, что отличает биологическую систему от простого набора компонентов. Невозможно изучать живое, рассматривая только материальные компоненты системы и упуская ее функциональный компонент. Принятие этой идеи означает, что признается онтологический статус чего-то иного, чем только атомы и молекулы [18].

3. Сознание у животных

Вернемся к проблеме сознания. Сначала обсудим возможности движения по первому направлению, приведенному во введе-

нии: перенос фокуса внимания на более простые системы, обладающие сознанием.

Все больше исследователей [19-25] считают, что сознанием обладают все животные, способные к сложному поведению, поскольку такое поведение невозможно без наличия у животных внутренней репрезентации внешнего окружения.

Одним из ярких примеров является прохождение рыбками зеркального теста Гэллапа [26], считающегося критерием наличия самосознания и заставляющего либо признать наличие у этих рыбок (с мозгом, содержащим 5 млн нейронов) сознания, либо отказаться от признанного теста.

Эксперименты с общественными насекомыми (муравьи, пчелы, шмели), имеющими не более 1 млн нейронов [27-31], продемонстрировали, что эти организмы обладают внутренней картиной мира, а также способны к категоризации, абстрагированию и даже к манипулированию позициями восприятия.

Из вышеизложенного можно сделать вывод, что нужно перестать воспринимать самосознание животных «как нечто черно-белое», как состояние «нет» либо «есть». Эта функция мозга может эволюционно развиваться последовательно, через серию усложняющихся стадий. Очевидное существование градаций уровня сознания и/или осознавания открывает перспективы исследования феноменов сознания на более простых организмах, а значит, и их моделирование с помощью существенно более простых моделей.

4. Рефлексия – ключевая характеристика сознания

Теперь, в соответствии с подходом биофизики, соответствующим второму направлению движения, приведенному во введении, и скорректированным в соответствии с идеями Розена, нужно выделить (фракционировать) ключевую функцию сознания и исследовать ее свойства на системах-аналогах.

Наличие внутренней репрезентации внешнего мира (рефлексия в широком смысле, понимаемая как отражение внешнего мира), по мнению многих специалистов, является ключевым свойством или характеристикой сознания [19-21]. А появление в этой репрезентации внешнего мира образа себя (появление третьей позиции восприятия, или рефлексии в узком смысле, как ее понимают психологи) некоторые авторы рассматривают как появление сознания, т.е., по их мнению, сознание = рефлексия [32; 33].

Тем самым мы подходим к наличию своеобразной линейки систем-аналогов: мозг человека, мозг позвоночных, мозг беспозвоночных, искусственные нейронные сети, т. е. эвристические модели по фон Нейману. Причем ключевое свойство или выделенная функция – рефлексия – может регистрироваться в эксперименте достаточно определенно.

Следуя логике эвристического моделирования для выделения сути рефлексивных процессов в узком смысле, нужно исследовать такие виды поведения, которые содержат минимальный вклад других когнитивных функций – логического рассуждения, распознавания образов, памяти и др. Этим требованиям почти идеально соответствуют рефлексивные игры [34; 35].

Еще одним перспективным видом экспериментов с эвристическими моделями рефлексии является постановка перед нейронной сетью такой задачи, которая не может быть решена без внутреннего отображения окружающего мира. Примером такой задачи является тест Отложенного Сравнения с Образцом (ОСО-тест), который был в частности использован для доказательства наличия рефлексии у пчел [36]. С него и начнем рассмотрение примеров использования эвристических нейросетевых модельных объектов.

Результаты исследования и обсуждение

1. Прохождение нейронной сетью теста Отложенного Сравнения с Образцом

В работах [37; 38] использовались простые рекуррентные нейронные сети (РНС), имеющие два входа, два выхода и содержащие 25 внутренних нейронов. Данное число нейронов было определено эмпирически как минимально необходимое для решения задачи.

Начальные значения весовых коэффициентов выбирались случайным образом из диапазона (–0.025; 0.025). Отклик РНС в момент времени t регистрировался на двух выходных нейронах:

$$y_h^t = f_h(W_h \cdot y_h^{(t-1)} + W_i \cdot x^{(t)}),$$
$$y_o^t = f_o(W_o \cdot y_h^{(t)}). \quad (1)$$

где Wh, Wi, Wo – матрицы весовых коэффициентов внутренних нейронов, входов и выходных нейронов соответственно; $x^{(t)}$ – вектор входных сигналов в момент времени t; $y^{(t)}$ и $y^{(t-1)}$ – векторы, описывающие уровни возбуждения внутренних нейронов в мо-

менты времени t и t-1, $f_h(.)$ и $f_o(.)$ – функции активации внутренних и выходных нейронов соответственно. Для простоты в уравнениях опущены смещения нейронов.

Активационная функция внутренних нейронов имела сигмоидный вид (2a). Кусочно-линейная функция активации (2b) выходных нейронов использовалась для получения точного выходного сигнала 0/1.

$$a) f_h(x) = \frac{1}{2}\left(\frac{x}{a+|x|}+1\right),$$
$$b) f_o(x) = \begin{cases} 0, if\ x \le 0, \\ b \cdot x, if\ 0 < x < 1, \\ 1, if\ x \ge 1. \end{cases} \quad (2)$$

Параметры функций активации (2) имели значения a=0,1 и b=1, подобранные эмпирически для наиболее быстрого обучения РНС. Шаг модификации синапсов задавался равным 10^{-3}.

Обучение РНС проводилось с помощью алгоритма обратного распространения ошибки. Здесь и во втором примере использовалась квадратичная функция потерь:

$$C = \frac{1}{2}\sum_{i}^{N}\left(\alpha_i^t - \delta_i^t\right)^2 \quad (3)$$

где α_i^n – сигнал на выходных нейронах в момент времени n, δ_i^n – *требуемый от сети сигнал в момент времени* t, N – количество выходных нейронов.

В настоящей работе отложенный тест сравнения проводился следующим образом. На вход РНС в случайный момент времени поступал один из трех случайно выбранных стимулов – входных векторов: A – (01), B – (10) и C – (11). Далее наступала пауза, продолжительность которой выбиралась случайно в интервале от 3 до 6 тактов. Затем на вход РНС поступал второй стимул из трех возможных, также выбранный случайно. На третий такт после предоставления второго стимула РНС должна была выдать одиночный сигнал на первом выходном нейроне, если стимулы совпадали, и одиночный сигнал на втором выходном нейроне, если стимулы различались (рисунок 1).

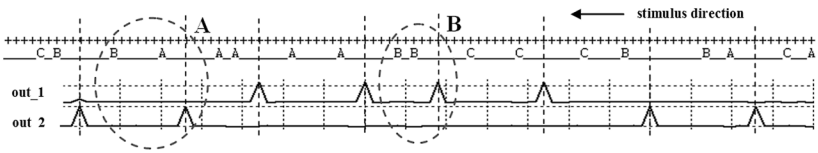

Рисунок 1. Фрагмент последовательности входных сигналов и откликов РНС. Знаки (+) показывают качество функционирования. Строка (A, B, C,_) показывает входной поток стимулов. Out_1 и out_2 соответствуют сигналам выходных нейронов. Пунктирной линией обведены примеры отклика РНС на различающиеся (A) и совпадающие (B) стимулы

Оказалось, что информация о первом стимуле хранится в РНС в виде динамического паттерна, причем различия между паттернами активности, соответствующими данному стимулу в разные моменты времени, более значительны, чем между паттернами, соответствующими разным стимулам в один и тот же момент времени (рисунок 2).

Для выявления общих закономерностей были использованы различные методы – от простого сравнения средних и метода центроидов [39] до метода кросс-временной классификации [40]. Кросс-временная классификация показала наличие возможности для декодирования внутренних динамических паттернов нейронной активности.

Попытка осуществить индентификацию обрабатываемого в данный момент стимула по паттерну нейронной активности методом центроидов показала, что его эффективность не превышает 80%, что объясняется уже упомянутой высокой вариабельностью сигнала.

Надежное распознавание содержания перерабатываемой нейросетью информации по динамическим паттернам возбуждений нейронов было достигнуто при использовании нейронной сети, выполняющей роль Нейросетевого Детектора (НД). В качестве НД использовалась однослойная нейронная сеть из трех нейронов с линейной характеристикой (2b). Каждый нейрон имел модифицируемый синапс с каждым из входов, число которых равнялось количеству нейронов РНС. НД выдавал единицу на одном из трех

нейронов, соответствующем приписанному стимулу, и нули на остальных. Для обучения использовался алгоритм обратного распространения ошибки. Обученные НД декодировали репрезентируемые РНС стимулы с точностью 100%.

Тем самым однослойная нейронная сеть оказалась способна надежно определять по динамически изменяющемуся паттерну активности нейронов вид инициирующего эту активность стимула. Детекторная нейронная сеть фактически выделила линейный инвариант возбуждений каждого из стимулов. Дополнительная процедура редукции сложности нейронной сети позволяет получать минимальный по сложности линейный инвариант. На рисунке 2 в качестве иллюстрации показана динамика внутреннего возбуждения нейронов, выделенных нейронным детектором, для стимулов *A, B* и *C*.

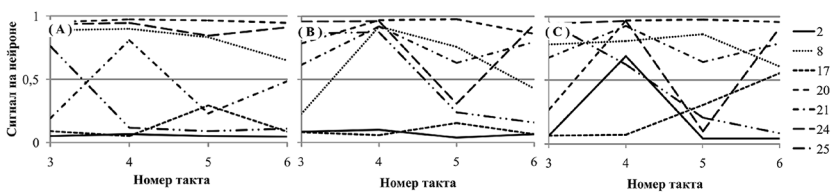

Рисунок 2. Динамика возбуждения нейронов, выделенных нейронным детектором, репрезентирующих соответствующий входной стимул (*A*, *B* и *C*)

Нужно отметить, что для каждой из РНС, проходящих ОСО-тест, нужно формировать свой нейросетевой детектор, т.е. у каждой нейронной сети свой «индивидуальный» нейродинамический код рефлексивного представления внешнего воздействия. Однако сложности найденных линейных инвариантов для разных нейросетей, проходящих один и тот же ОСО-тест, оказались сопоставимы, что, впрочем, и следовало ожидать.

2. Сравнение эффективности рекуррентной и многослойной нейронных сетей: рефлексия против регрессии

В работе [41] были использованы полносвязные РНС и многослойные нейронные сети (МНС) из трех слоев с разным количеством нейронов. Кроме того, мы использовали различную глубину распространения ошибок для РНС и длину регистра сдвига, отображающего предыдущие ходы партнера по игре для МНС.

Нейросети играли друг с другом в рефлексивную игру «чет-нечет», в которой нет фиксированной выигрышной стратегии, а ничейный в среднем результат достигается случайным выбором хода. Если игроки пытаются избежать ничьей, то игра становится нетривиальной: победителем становится тот, кто лучше предсказывает ходы противника.

Чтобы выделить только влияние структуры на игровые способности нейросетей, функции перехода для РНС и МНС были выбраны идентичными:

$$\alpha_i^{n+1} = \frac{\rho_i^n}{a + \left|\rho_i^n\right|}, \rho_i^n = \sum_j w_{ij}\alpha_j^n + A_i^n, \quad (4)$$

где w_{ij} – матрица весовых коэффициентов, A_i^n – входные сигналы, α_i^n – выходной сигнал j-го нейрона в n-ый момент времени, a – константа, задающая крутизну активационной функции нейрона.

Информация о ходах партнера подается через два входных нейрона РНС и в первый разряд сдвигового регистра МНС. Соотношение сигналов двух выходных нейронов определяет ход нейронной сети – “0” или “1”. Целевые функции для нейросетей различаются, потому что одна нейросеть выигрывает, когда ее ход совпадает с ходом второй нейросети, а вторая нейросеть выигрывает, когда делает ход, отличный от хода противника:

$$\begin{aligned} H_1(\alpha_i^n, n) &= \frac{1}{2}\left[\left(\alpha_3^n - move2\right)^2 + \left(\alpha_4^n - (1 - move2)\right)^2\right], \\ H_2(\alpha_i^n, n) &= \frac{1}{2}\left[\left(\alpha_3^n - (1 - move1)\right)^2 + \left(\alpha_4^n - move1\right)^2\right]. \end{aligned} \quad (5)$$

Синапсы РНС и МНС изменялись после каждого хода в соответствии с алгоритмом обратного распространения ошибки.

Сравнение игровых способностей РНС и МНС показало, что при одинаковом количестве нейронов и глубине памяти (глубина распространения ошибки в прошлое) МНС в целом демонстрируют значительно худшее качество игры, чем РНС (МНС выигрывала в среднем в ~10% случаев).

В рефлексивных играх устойчивый выигрыш подразумевает, что у выигравшего игрока рефлексия на один ранг выше, чем у соперника. В пользу предположения о том, что в большинстве проанализированных игровых паттернов (рисунок 3) выигрыш был не результатом случайного успеха, а результатом «сознательного»

(рефлексивного) выбора хода, говорит разделение всех игровых паттернов на две группы (рисунок 4). В случае случайных исходов игры такая кластеризация представляется невозможной.

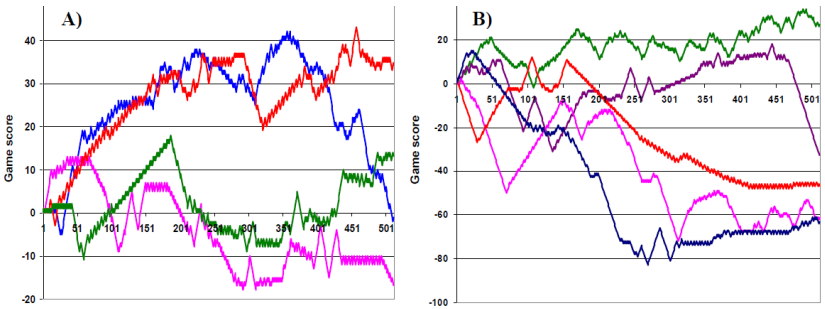

Рисунок 3. Примеры игровых паттернов двух РНС (А) и рекуррентных и многослойных нейросетей (В)

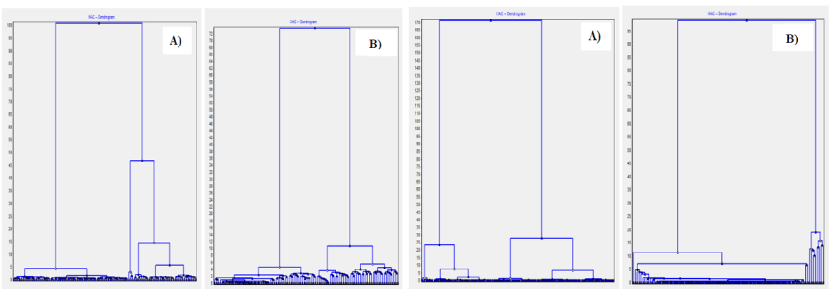

Рисунок 4. Иерархическое дерево близости частотных (А) и фазовых (В) спектров преобразования Фурье игровых паттернов между рекуррентными (левая пара) и рекуррентными и многослойными (правая пара) нейросетями

Тем самым было продемонстрировано, что регрессия не может эффективно заменить рефлексию в некооперативных стратегических взаимодействиях, которые включают рефлексивные игры. Из чего следует, что разработка по-настоящему мощного искусственного интеллекта невозможна без воспроизведения рефлексивных процессов в интеллектуальных системах.

3. О нейрофизиологической природе асимметрии игры «Чет-нечет»

Вопрос о возможной асимметрии игры «чет-нечет» был поднят в статье [42] на примере мальчика из рассказа Э. А. По «Похищенное письмо». Этот мальчик оценивал ум и сообразительность противника и в зависимости от этого делал свои ходы. Авторы отметили, что мальчик всегда выбирал позицию угадывающего, т. е. он работал за «чет» – получал выигрыш, если стороны монеток совпадали. Авторы статьи задались вопросом о возможной асимметрии игры и предположили, что игрок за «чет» имеет какое-то преимущество. Они провели эксперименты с 55 парами игроков, которые играли в «чет-нечет» длительностью 24 раунда.

Было показано, что, действительно, игроки, играющие за «чет», выигрывали в среднем в 54% случаев, причем статистические оценки показали, что это отличие является достоверным. Авторы проверяли в эксперименте различные причины асимметрии, включая психологические, задавая различные формулировки ролей участников, например *«вводящий в заблуждение-угадывающий»* или *нейтральные «четная и нечетная игра»* или *«делающий первый ход и второй».*

Среди гипотез была также чисто нейрофизиологическая или информационная, что задача игрока за «нечет» сложнее – нужно не только предсказать ход противника, но и выбрать противоположный ход для себя.

В конечном счете, авторы отметили, что могут одновременно срабатывать различные причины, подчеркнув, что даже в такой простой и знакомой игре, с очень простыми стратегиями равновесия, наблюдается систематическое отклонение от предсказания теории игр.

Для проверки вклада чисто информационной причины асимметрии нами были проведены игры совершенно идентичных нейронных сетей с генерируемыми случайным образом синаптическими весами. Было проведено более 10000 тыс. игр с нейронными сетями с разным числом нейронов – 10, 15 и 20. В среднем асимметрия выигрышей была ~54% в пользу «чет», т. е. удивительным образом совпала со значением, полученным в рассматриваемой статье.

Поскольку психологическая причина асимметрии в данном случае отсутствует полностью, то приходится признать, что пре-

имущество игрока, играющего за «чет», определяется в первую очередь различием в сложности принятия решения об очередном ходе. Хотя, конечно, полученный результат не отменяет и психологические причины асимметрии исходов в игре «чет-нечет», когда в нее играют люди. Кстати, подобные эксперименты с нейросетями, играющими в «камень-ножницы-бумага», показали отсутствие какой-либо асимметрии, что, впрочем, и ожидалось.

Заключение

На основании вышеописанного можно заключить, что использование простых систем-аналогов данного динамического свойства или данной функциональной характеристики системы – в данном случае это рефлексия – позволяет получить результаты, которые трудно или просто невозможно получить в реальном эксперименте. Например, очень трудно заставить 10000 человек играть в рефлексивную игру из 500 раундов, или получить доступ к каждому нейрону ансамбля, осуществляющего динамическую репрезентацию поступившего стимула, или отключить у испытуемого рефлексию и оставить только адаптивную рефлекторную реакцию.

Кроме того, эвристический модельный объект, в качестве которого в работе выступали рекуррентные нейронные сети, «прозрачен», про его структуру известно все и это позволяет сконцентрировать усилия непосредственно на выявлении общих закономерностей, а не на поиске кажущихся важными, а по сути бесконечных, деталей исследуемой структуры.

При этом нужно отметить своеобразный парадокс: несмотря на то что в экспериментах с эвристическими нейросетевыми моделями используются очень простые абстрактные стимулы и сама нейронная сеть является в высшей степени абстрактной моделью реальной биологической системы, она позволяет достигнуть конкретности в описании столь трудно формализуемых свойств, как «структура», «функция», «сложность» и «рефлексия», что очень трудно сделать в отношении живых систем.

Представляется, что дальнейшее использование данного подхода будет продуктивным и позволит найти ответы на вопросы, которые легче поставить, чем получить на них ответ.

Литература

1. *Seth A. K., Bayne T.* Theories of consciousness // Nature Reviews Neuroscience. – 2022. – Vol. 23. – № 7. – P. 439-452.

2. *Васильев В. В.* Трудная проблема сознания. – М.: Прогресс-Традиция, 2009. – 272 с. (*Vasil'ev V. V.* Hard problem of consciousness. – M.: Progress-Traditsiya, 2009. – 272 p.)
3. *Ревонсуо А.* Психология сознания. – СПб.: Питер, 2013. – 309 с. (*Revonsuo A.* Psychology of consciousness. – SPb.: Piter, 2013. – 309 p.)
4. *Чалмерс Д.* Сознающий ум. В поисках фундаментальной теории. – М.: УРСС: Книжный дом «ЛИБРОКОМ», 2003. – 512 с. (*Chalmers D.* Conscious mind. In search of a fundamental theory. – M.: URSS: Knizhniy dom «LIBROKOM», 2003. – 512 p.)
5. *Пенроуз Р.* Тени разума: в поисках науки о сознании. – М.-Ижевск: Институт космических исследований, 2005. – 688 с. (*Penrose R.* Shadows of the Mind: In Search of a Science of Consciousness // R. Penrose. – M.-Izhevsk: Institut kosmicheskih issledovanij, 2005. – 688 p.)
6. *Хренников А. Ю.* Моделирование процессов мышления в p-адических системах координат. – М.: ФИЗМАТЛИТ, 2004. – 296 с. (*Khrennikov A. Yu.* Modeling of thinking processes in p-adic coordinate systems. – M.: FIZMATLIT, 2004. – 296 p.)
7. *Crick F., Koch C.* Towards a neurobiological theory of consciousness // Seminars in the Neurosciences // Saunders Scientific Publications. – 1990. – Vol. 2. – P. 263-275.
8. *Crick F., Koch C.* A framework for consciousness // Nature neuroscience. – 2003. – Vol. 6. – № 2. – P. 119-126.
9. *Frith C.* The quest for consciousness: A neurobiological approach // American Journal of Psychiatry. – 2005. – Vol. 162. – № 2. – P. 407-407.
10. *Барцев С. И., Барцева О. Д.* Эвристические нейросетевые модели в биофизике: приложение к проблеме структурно-функционального соответствия. – Красноярск: Сибирский федеральный университет, 2010. – 115 с. (*Bartsev S. I., Bartseva O. D.* Heuristic neural network models in biophysics: application to the problem of structural-functional correspondence. – Krasnoyarsk: Sibirskij federal'nij universitet, 2010. – 115 p.)
11. *Блюменфельд Л. А.* Решаемые и нерешаемые проблемы биологической физики. – М.: Едиториал УРСС, 2002. – 160 с. (*Blumenfeld L. A.* Solvable and unsolvable problems of biological physics. – M.: Editorial URSS, 2002. – 160 p.)
12. *Моровиц Г.* Исторический очерк // Теоретическая и математическая биология. – М.: Мир, 1968. – С. 34-48. (*Morovits G.* Historical sketch // Teoreticheskaya and matematicheskaya biologiya. – M.: Mir, 1968. – P. 34-48.)
13. *Фон Нейман Дж.* Теория самовоспроизводящихся автоматов. – *М.*: *Мир*, 1971. – *С.* 382. (*Von Neumann J.* Theory of self-reproducing automata. – M.: Mir, 1971. – P. 382.)
14. *Бернал Дж. Д.* Молекулярная структура, биохимическая функция и эволюция // Теоретическая и математическая биология. – М.: Мир, 1968. – С. 110-151. (*Bernal J. D.* Molecular structure, biochemical physics and evolution / Teoreticheskaya i matematicheskaya biologiya. – M. Mir, 1968. – P. 110-151.)

15. *Рашевский Н.* Модели и математические принципы в биологии // Теоретическая и математическая биология. – М.: Мир, 1968. – 448 с. (*Rashevsky N.* Models and mathematical principles in biology // Teoreticheskaya and matematicheskaya biologiya. – M.: Mir, 1968. – 448 p.)
16. *Rosen R.* A relational theory of biological systems // The bulletin of mathematical biophysics. – 1959. – Vol. 21. – P. 109-128.
17. *Lennox J.* Robert Rosen and relational system theory: an overview // PhD Dissertation. – The City University of New York, 2022. – 195 p.
18. *Mikulecky D. C.* Complexity, communication between cells, and identifying the functional components of living systems: some observations // Acta Biotheoretica. – 1996. – Vol. 44. – № 3-4. – P. 179-208.
19. *Bickhard M. H.* Consciousness and reflective consciousness // Philosophical Psychology. – 2005. – Vol. 18. – № 2. – P. 205-218.
20. *Dehaene S., Lau H., Kouider S.* What is consciousness, and could machines have it? // Science. – 2017. – Vol. 358. – № 6362. – P. 486-492.
21. *Land M. F.* Do we have an internal model of the outside world? // Philosophical Transactions of the Royal Society B: Biological Sciences. – 2014. – Vol. 369. – № 1636. – P. 20130045.
22. *Chang A. Y. C., Biehl M., Yu Y., Kanai R.* Information closure theory of consciousness // Frontiers in Psychology. – 2020. – Vol. 11. – P. 1504.
23. *Lamme V. A. F.* Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism // Philosophical Transactions of the Royal Society B: Biological Sciences. – 2018. – Vol. 373. – № 1755. – P. 20170344.
24. *Zalucki O., Van Swinderen B.* What is unconsciousness in a fly or a worm? A review of general anesthesia in different animal models // Consciousness and cognition. – 2016. – Vol. 44. – P. 72-88.
25. *Nieder A., Wagener L., Rinnert P.* A neural correlate of sensory consciousness in a corvid bird // Science. – 2020. – Vol. 369. – № 6511. – P. 1626-1629.
26. *Kohda M. et al.* Further evidence for the capacity of mirror self-recognition in cleaner fish and the significance of ecologically relevant marks // PLoS biology. – 2022. – Vol. 20. – № 2. – P. e3001529.
27. *Alem S., Perry C. J., Zhu X., Loukola O. J., Ingraham T., Søvik E., Chittka L.* Associative mechanisms allow for social learning and cultural transmission of string pulling in an insect // PLoS biology. – 2016. – Vol. 14. – № . 10. – P. e1002564.
28. *Avarguès-Weber A., Giurfa M.* Conceptual learning by miniature brains // Proceedings of the Royal Society B: Biological Sciences. – 2013. – Vol. 280. – № 1772. – P. 20131907.
29. *Howard S. R., Avarguès-Weber A., Garcia J. E., Greentree A. D., Dyer A. G.* Numerical ordering of zero in honey bees // Science. – 2018. – Vol. 360. – № 6393. – P. 1124-1126.
30. *Loukola O. J., Perry C. J., Coscos L., Chittka L.* Bumblebees show cognitive flexibility by improving on an observed complex behavior //

Science. – 2017. – Vol. 355. – № 6327. – P. 833-836.
31. *Ulrich Y., Saragosti J., Tokita C. K., Tarnita C. E., Kronauer D. J. C.* Fitness benefits and emergent division of labour at the onset of group living // Nature. – 2018. – Vol. 560. – № 7720. – P. 635-638.
32. *Лефевр В. А.* Рефлексия. – М.: Когито-Центр, 2003. – 496 с. (*Lefebvre V. A.* Reflection. – M.: Kogito-Tsentr, 2003. – 496 p.)
33. *Peters F.* Theories of consciousness as reflexivity // The Philosophical Forum. – 2013. – Vol. 44. – P. 341-372.
34. *Лефевр В. А.* Лекции по теории рефлексивных игр. – М.: Когито-Центр, 2009. – 218 с. (*Lefebvre V. A.* Lectures on the theory of reflexive games. – M.: Kogito-Tsentr, 2009. – 218 p.)
35. *Camerer C. F., Ho T. H., Chong J. K.* A cognitive hierarchy model of games // The Quarterly Journal of Economics. – 2004. – Vol. 119. – № 3. – P. 861-898.
36. *Giurfa M.* Behavioral and neural analysis of associative learning in the honeybee: a taste from the magic well // Journal of comparative physiology A. – 2007. – Vol. 193. – № 8. – P. 801-824.
37. *Барцев С. И., Батурина П. М., Маркова Г. М.* Нейросетевое декодирование информации о внешнем стимуле по паттерну нейронной активности рекуррентной нейронной сети // Доклады Российской академии наук. Науки о жизни. – 2022. – Т. 502. – № 1. – С. 48-53. (*Bartsev S. I., Baturina P. M., Markova G. M.* Neural network-based decoding input stimulus data based on recurrent neural network neural activity pattern // Doklady Biological Sciences. – M.: Pleiades Publishing, 2022. – Vol. 502. – № 1. – P. 1-5.)
38. *Bartsev S. I., Markova G. M.* Decoding of stimuli time series by neural activity patterns of recurrent neural network // Journal of Physics: Conference Series. – IOP Publishing, 2022. – Vol. 2388. – № 1. – P. 012052.
39. *Crowe D. A., Averbeck B. B., Chafee M. V.* Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex // Journal of Neuroscience. – 2010. – Vol. 30. – № 35. – P. 11640-11653.
40. *Meyers E. M., Freedman D. J., Kreiman G., Miller E. K., Poggio T.* Dynamic population coding of category information in inferior temporal and prefrontal cortex // Journal of neurophysiology. – 2008. – Vol. 100. – № 3. – P. 1407-1419.
41. *Bartsev S., Markova G.* Recurrent and multi-layer neural networks playing Even-Odd: reflection against regression // IOP Conference Series: Materials Science and Engineering. – IOP Publishing, 2020. – Vol. 734. – № 1. – P. 012109.
42. *Eliaz K., Rubinstein A.* Edgar Allan Poe's riddle: Framing effects in repeated matching pennies games // Games and Economic Behavior. – 2011. – Vol. 71. – № 1. – P. 88-99.