Методологические аспекты информационных и математических технологий

УДК 577.38+004.81
DOI:10.25729/ESI.2025.37.1.001

Предельно просто не значит предельно ясно: некоторые контринтуитивные результаты нейросетевого моделирования рефлексии

Маркова Галия Муратовна[1,2], Барцев Сергей Игоревич[1,2]

[1]Институт биофизики СО РАН – обособленное подразделение ФИЦ КНЦ СО РАН, Россия, Красноярск, *GMarkova@ibp.ru*

[2]Сибирский федеральный университет, Институт фундаментальной биологии и биотехнологий, Россия, Красноярск

Аннотация. В работе представлен ряд результатов по моделированию рефлексии, понимаемой в широком смысле, как наличие у активного агента внутреннего отображения внешнего мира, влияющего на его поведение. Выявлена способность простейших нейросетевых модельных объектов гомогенной и гетерогенной (модульной) структуры к решению задач, требующих наличия и использования устойчивых внутренних отображений (репрезентаций) внешних стимулов. Определено, что данные репрезентации являются декодируемыми, т.е. по текущему виду паттерна нейронной активности нейросетевого объекта возможно определить, какой конкретно стимул или временной ряд стимулов в данный момент в нём обрабатывается. Приведены изначальные предположения авторов, сделанные из общих соображений относительно эффективности нейросетевых модельных объектов различной структуры в задачах на рефлексию, и соответствующие полученные результаты. В частности, показаны следующие эффекты: 1) позиции в игре чет-нечет асимметричны при условии ограниченности вычислительных возможностей игроков; 2) формально близкие задачи на рефлексию (игра чет-нечет и реагирование на фиксированные временные ряды стимулов по правилам этой игры) различаются по требованиям к игрокам; 3) декодируемыми являются паттерны нейронной активности не только нейронных сетей, обученных реагированию на стимулы, но и сетей со случайными весовыми коэффициентами; 4) точность декодирования нейронной активности рекуррентных нейронных сетей, обладающих гетерогенностью во времени, превосходит точность отклика этих сетей при реагировании на ряды стимулов; 5) паттерны нейронной активности у гомогенных рекуррентных нейронных сетей сложнее для декодирования, чем у гетерогенных сопоставимого размера. Данные эффекты иллюстрируют богатую внутреннюю и поведенческую динамику простейших рекуррентных нейронных сетей, что, с одной стороны, перспективно для исследовательских и практических целей, а с другой – затрудняет предсказание и интерпретацию поведения объектов подобного рода.

Ключевые слова: рекуррентные нейронные сети, рефлексия, рефлексивные игры, декодирование нейронной активности, репрезентация внешних стимулов

Цитирование: Маркова Г.М. Предельно просто не значит предельно ясно: некоторые контринтуитивные результаты нейросетевого моделирования рефлексии / Г.М. Маркова, С.И. Барцев // Информационные и математические технологии в науке и управлении, 2025. – № 1(37). – С. 5-15. – DOI:10.25729/ESI.2025.37.1.001.

Введение. Под рефлексией (в широком смысле) понимается феномен наличия у активного агента внутреннего представления внешнего мира, которое влияет на его активность. Условие влияния на активность является ключевым, поскольку в противном случае отпечаток ступни на мокром песке тоже можно считать отображением, что делает такое понимание отображения внешнего мира тривиальным и неоперациональным. Сохранение в памяти и своевременное извлечение из нее адекватных представлений о мире дает возможность осуществлять прогностическую обработку поступающей информации [1, 2]. Такое опережающее отражение действительности [3] позволяет реализовывать наиболее эффективное поведение, поскольку действующий агент принимает решение относительно ожидаемых событий и подготовлен к наиболее адекватному действию. В этом состоит отличие поведения, основанного на рефлексии, от реактивного или «рефлекторного» поведения.

Рефлексивное поведение демонстрируют не только люди, но и многие животные, в том числе, с чрезвычайно малым мозгом по сравнению с человеческим, например, шмели [4]. В таком случае естественно предположить, что поведение, основанное на рефлексии, может быть воспроизведено и исследовано на простых объектах, причем необязательно биологических, а, например, на искусственных нейронных сетях. Подобные модельные объекты устроены существенно проще своих биологических прототипов, так что исследователи получают полный контроль над внешними воздействиями на объект и могут отслеживать все его внутренние состояния. То, что простые нейронные сети, состоящие из 15-30 формальных нейронов, способны проходить классические тесты на рефлексию, например, тест отложенного сравнения с образцом, показано нами ранее [5].

Согласно биофизическому подходу к моделированию рефлексии [6], в работе использовались наиболее простые конфигурации нейронных сетей, функционирующих в дискретном времени: полносвязные рекуррентные нейронные сети (РНС) без отдельных слоев для входа и выхода. Рекуррентная структура вместо прямого распространения выбрана в связи с тем, что динамическое отображение внешнего мира реализуемо только при условии использования рабочей памяти – кратковременного хранения информации, позволяющего судить о контексте, в котором появляются новые стимулы. Аналогом рабочей памяти в нейронной сети может служить рекуррентное возбуждение нейронов.

Функционирование РНС задавалось формулами:

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

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. Стимулы «0» и «1» кодировались как пары сигналов (01) и (10) соответственно, поэтому поступали на два входа. Отклик РНС определялся по соотношению сигналов двух выходных нейронов. Обучение РНС проводилось по классическому алгоритму backpropagation с глубиной распространения ошибки 5, использовалась квадратичная функция потерь.

В ряде экспериментов было обнаружено, что, несмотря на сравнительную простоту устройства, данные модельные объекты демонстрируют труднопредсказуемое, контринтуитивное поведение при решении простейших задач на рефлексию. Статья посвящена рассмотрению некоторых обнаруженных неожиданностей.

1. Асимметрия позиций в рефлексивной игре чет-нечет. По литературным данным [7], в рефлексивной игре чет-нечет наблюдается смещение побед в пользу игрока за позицию «чет», т.е. того, кто должен предсказать ход противника на каждом шаге игры и сделать такой же. Это смещение связано с фреймингом – особенностями восприятия своей позиции у игроков. Игроки, не обладающие сложной психикой (например, РНС), не подвержены фреймингу. *Было сделано предположение,* что смещение в пользу РНС-игрока за «чет» всё равно может присутствовать, поскольку его задача – только предсказать ход противника и воспроизвести его, в то время как игрок за «нечет» должен предсказать и «перевернуть» этот ход, чтобы самому сделать противоположный, что является более сложной вычислительной процедурой.

Были использованы РНС, функционирующие согласно уравнениям (1), в нескольких вариантах модификаций: 1) рекуррентная полносвязная сеть, без отдельных слоев входа и выхода, размер 15 нейронов (SRN15); 2) аналогичная предыдущей, но с дополнительным входом, куда в виде сигнала +1/-1 поступают сведения о победе/поражении на предыдущем шаге игры (SRN+15); 3) аналогичная первой, но из 30 нейронов (SRN30); 4) аналогичная второй, но из 30 нейронов (SRN+30).

В эксперименте РНС играли против себе подобных. Весовые коэффициенты перед началом партии задавались случайным образом в диапазоне (-0.025; 0.025) и модифицировались после каждого хода по алгоритму backpropagation с шагом 0,003. В партиях длиной 1000 ходов регистрировались показатели успешности РНС-игроков: 1) количество баллов, набранных РНС при игре за «чет» и за «нечет», 1500 партий за каждую позицию; 2) количество побед РНС при игре за «чет» и за «нечет», 15 раз по 100 партий за каждую позицию. Сравнивались средние значения показателей каждой конфигурации для разных позиций с помощью двухвыборочного *t*-теста с различными дисперсиями. Результаты показаны в таблице 1, в формате (среднее ± ошибка среднего). Статистически значимые различия между средними значениями обоих показателей были зарегистрированы только для РНС типа SRN15, т.е. сетей минимального размера и с простейшей структурой.

Таблица 1. Показатели успешности РНС-игроков за позиции «чет» и «нечет»

	Средняя доля выигрышных ходов за партию, из 1000 ходов ($t_{кр}$ = 1,96)			Среднее количество побед, из 100 партий ($t_{кр}$ = 2,02)		
Тип сети	Нечет	Чет	t-статистика	Нечет	Чет	t-статистика
SRN15	(49,7 ± 0,2)%	(50,2 ± 0,2)%	-4,6	48 ± 2	51 ± 2	-2,1
SRN+15	(49,9 ± 0,1)%	(50,0 ± 0,1)%	-0,5	49 ± 3	50 ± 2	-0,5
SRN30	(50,0 ± 0,1)%	(50,1 ± 0,1)%	-1,5	49 ± 3	50 ± 3	-0,7
SRN+30	(49,9 ± 0,1)%	(49,9 ± 0,1)%	-1,2	49 ± 3	48 ± 3	0,3

Следовательно, асимметрия позиций в игре чет-нечет наблюдается, если вычислительные возможности игрока ограничены. Однако даже небольшая модификация структуры РНС (введение дополнительного входа или увеличение количества нейронов) позволяет устранить этот эффект или, по крайней мере, сделать его малозаметным.

2. Эффективность гомогенных и гетерогенных РНС в разных задачах. Помимо игры чет-нечет, рассматривалась задача реагирования РНС на фиксированные временные ряды стимулов, сменявшие друг друга случайным образом. Эта задача соответствует смене ситуаций, происходящих при активном взаимодействии животного с окружающей средой. В качестве примера можно назвать охоту хищника, когда стадия поиска добычи или ожидания в засаде сменяется стадией активного преследования, но момент появления добычи заранее не известен. Более того, может появиться конкурент или более сильный хищник, и тогда необходима другая программа действий. В данной работе смена ситуаций имитировалась подачей на вход РНС одного из четырех временных рядов (см. таблицу 2), сменявших друг друга в случайном порядке по истечении определенного количества игровых шагов. РНС нужно было как можно быстрее распознать ряд, который начал подаваться, и сформировать отклик. Правильность отклика определялась также по правилам игры чет-нечет, то есть от РНС требовалось предсказать следующий элемент ряда и сгенерировать либо совпадающий сигнал (как при игре за «чет»), либо противоположный (как при игре за «нечет»).

Таблица 2. Фиксированные последовательности стимулов – временные ряды

Ряд 1	Ряд 2	Ряд 3	Ряд 4
110011001100	101100101100	010011010011	111000111000

В данном эксперименте использовались РНС типа SRN30 и SRN+30, состоявшие из одного функционального модуля, т.е. гомогенные (см. рис. 1А), а также другие модификации РНС – гетерогенные, состоявшие из двух функционально различающихся модулей. Гетерогенные РНС типа «сети двойного времени» (Dual-time RNN, DTRNN) содержали медленный модуль, функционировавший по внешним тактам подачи стимулов, и быстрый модуль, имевший 3 внутренних такта на обработку стимула перед выдачей отклика (см. рис.

1Б). Гетерогенные РНС типа «рефлексивные сети» (Reflexive Net, RefNet) содержали играющий модуль, функционировавший, как SRN15, и рефлексивный модуль, который получал на вход как внешние стимулы, так и отклик играющего модуля, и мог изменить его на противоположный (см. рис. 1В). Все модули гетерогенных РНС по структуре представляли собой SRN15 или SRN+15. *Было сделано предположение*, что гетерогенные РНС обладают преимуществом над гомогенными того же размера (30 нейронов).

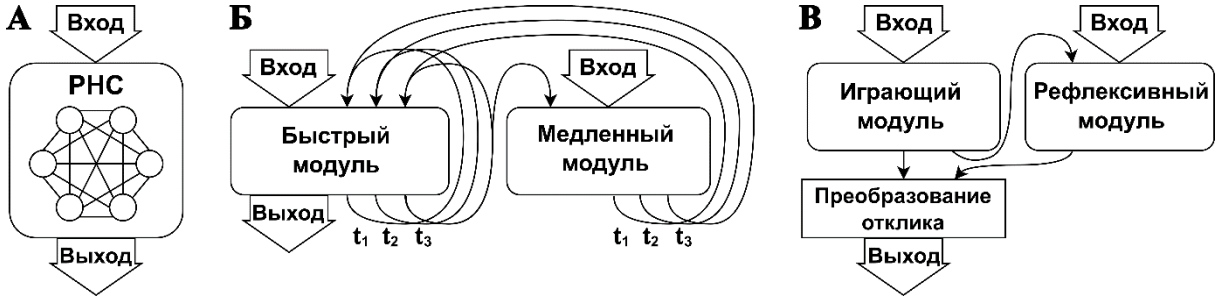

Рис. 1. Схемы структуры использованных РНС: А) SRN, SRN+ и все модули в составе гетерогенных РНС; Б) DTRNN, DTRNN+; В) RefNet, RefNet+

РНС, по 500 шт. каждой модификации, обучались в течение 1200 ходов. Ряды подавались по порядку, каждый в течение 60-и ходов (режим обучения). Далее весовые коэффициенты обученных РНС фиксировались, и на вход подавались те же ряды, каждый в течение 50-и ходов, в случайном порядке (режим теста). Для единообразия, и поскольку в предыдущем разделе статьи было показано, что успешность модифицированных РНС не зависит от позиции, все РНС функционировали в позиции «чет». Регистрировалась доля правильных откликов РНС в режимах обучения и теста. Важно отметить, что в данной задаче доля правильных ответов, равная 1, недостижима. После смены ряда РНС необходимо определить, что эта смена произошла, и распознать новый подаваемый ряд, поэтому в течение первых тактов подачи ряда (как правило, 3-5 тактов) РНС могут формировать неверный отклик.

На рис. 2А представлены результаты. Наилучшие показатели при обучении продемонстрировали гомогенные РНС: SRN (средняя доля правильных откликов 0,88) и SRN+ (0,83). Для них же характерно наибольшее снижение доли правильных ответов при тестировании. В режиме теста наибольшая средняя доля правильных ответов у гетерогенных РНС RefNet (0,81).

В целом, РНС с дополнительным входом в данной задаче показали результат хуже, чем без него. Однако, если сопоставлять эффективность тех же модификаций РНС с изменяемыми весовыми коэффициентами в игре чет-нечет против референсных РНС с гомогенной структурой (SRN), то наличие дополнительного входа, наоборот, способствует большему выигрышу (см. рис. 2Б). В этой задаче выигрыш всех гетерогенных РНС превысил таковой у контроля (SRN против SRN), $p < 0{,}001$ для всех конфигураций по t-тесту Уэлча. Но гомогенные РНС типа SRN+ также показали значимый выигрыш, превышающий таковой у гетерогенных РНС типа DTRNN и RefNet.

Следовательно, хотя обе рассмотренные задачи являются задачами на рефлексию (подразумевают наличие внутреннего представления о противнике – другой РНС в рефлексивной игре, или квазипротивнике – временных рядах стимулов), наиболее эффективными оказываются РНС с разными модификациями. Данный результат свидетельствует о многогранности феномена рефлексии.

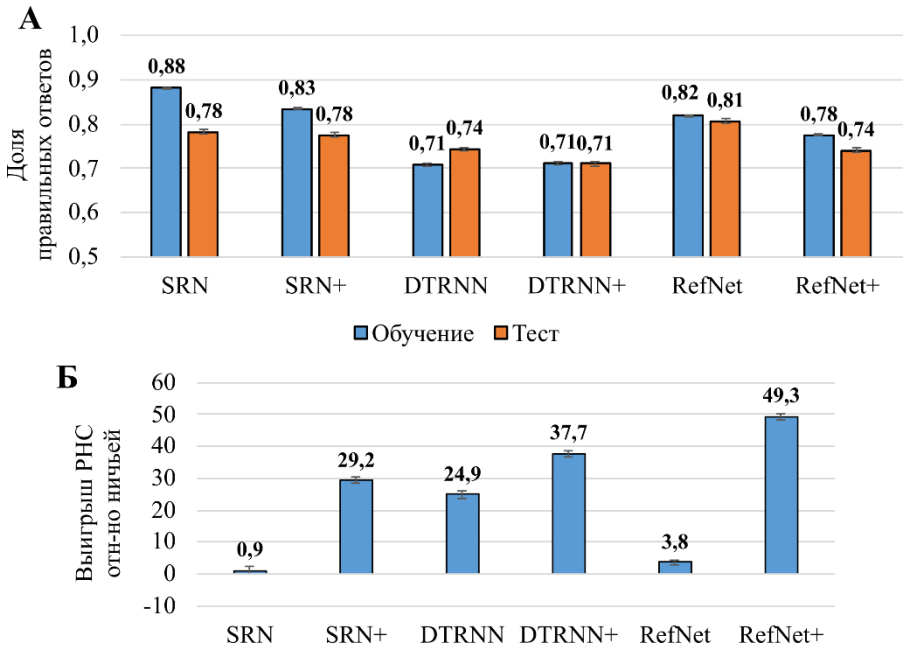

Рис. 2. А) эффективность РНС в реагировании на ряды; Б) эффективность РНС в игре чет-нечет против референсных РНС. В качестве погрешностей – ошибки среднего

3. Декодирование нейронной активности РНС, соответствующей различным стимулам. Внутренние представления внешнего мира, который для РНС сводится к определенным комбинациям получаемых стимулов, можно представить, как соответствующие этим стимулам паттерны нейронной активности. В этом случае возможно декодировать нейронную активность РНС, т.е. по виду паттерна распознать обрабатываемый в данный момент стимул.

3.1. Связь декодируемости нейронной активности и качества функционирования РНС. В вышеописанной задаче реагирования на фиксированные временные ряды стимулов, например, возможно распознать, какой ряд подается на РНС [8]. Было сделано предположение, что декодирование нейронной активности возможно только для РНС, хорошо освоивших задачу реагирования, т.к. формальный успех в данном случае свидетельствует о сформированности точных и правильных представлений о рядах.

РНС всех перечисленных ранее гомогенных и гетерогенных модификаций обучались задаче реагирования на фиксированные временные ряды стимулов в течение 1200 ходов. Были выбраны РНС каждой модификации, по 5 игравших за «чет» и за «нечет», показавшие наибольшую («лучшие») и наименьшую («худшие») доли правильных ответов в режиме теста. Также использовались РНС, получавшие в режиме обучения каждый ряд всего по одному разу в течение 24 шагов («one-shot») и не обучавшихся вообще («zero-shot»). Распознавание рядов по нейронной активности проводилось с помощью нейросетевого декодера (НС-декодера) – нейронной сети прямого распространения. Активационная функция внутренних нейронов НС-декодера имела сигмоидный вид (2a). Кусочно-линейная функция активации (2b) выходных нейронов НС-декодера использовалась для получения точного сигнала 0/1 на выходном нейроне, соответствующем номеру распознаваемого ряда:

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

Параметры функций активации (2) имели значения $a = 0.1$, $b = 1$, подобранные эмпирически для наиболее быстрого обучения НС-декодера, шаг модификации синапсов задавался равным 0,001. Декодирование производилось по отдельности для каждой РНС, т.к.

паттерны нейронной активности чрезвычайно вариативны. На вход декодеру подавалась нейронная активность РНС, записанная на одном такте при обработке текущего ряда. Требуемый отклик – сигнал, равный 1, на выходном нейроне, соответствующем номеру этого ряда, и 0 на остальных выходных нейронах.

Поскольку декодирование паттернов нейронной активности, в сущности, является задачей классификации, также использовался классический метод классификации k-ближайших соседей (KNN-декодер). Исходя из предварительных измерений, наилучший результат мог быть получен при $k = 3$. Помимо классов, соответствующих рядам стимулов, вводился класс «всё остальное». В обучении ему соответствовала нейронная активность РНС при подаче случайных последовательностей стимулов, не совпадавших с рядами. Предполагалось, что как «всё остальное» НС- и KNN-декодеры должны распознавать нейронную активность РНС, записанную на нескольких первых тактах при смене ряда, т.к. в это время происходит процесс переключения между репрезентациями рядов.

Декодирование нейронной активности, соответствовавшей четырем временным рядам стимулов, оказалось возможным для всех перечисленных выборок РНС («лучшие», «худшие», «one-shot», «zero-shot»), среднее качество декодирования около 0,8. У обученных РНС («лучшие» и «худшие») декодеры лучше выявляли класс «всё остальное» на первых тактах после смены ряда, подаваемого на вход РНС. Декодеры, распознававшие нейронную активность РНС из выборок «one-» и «zero-shot», в режиме теста не выделяли данный класс.

Данный результат соотносится с полученным ранее на другой задаче – тесте отложенного сравнения с образцом [9]: РНС формируют декодируемые динамические паттерны нейронной активности, соответствующие стимулам (или временным рядам стимулов), даже при отсутствии обучения, что обусловлено различиями в самих стимулах и внутренней нелинейной динамикой РНС. Благодаря обучению, в свою очередь, репрезентации стимулов становятся операциональными, т.е. РНС не просто отражают действительность, но и могут использовать это отражение для достижения цели. Сами репрезентации также становятся более оформленными, что позволяет декодерам уловить разницу между представлением рядов и промежуточными состояниями при переключении.

3.2. Декодирование нейронной активности гомогенных и гетерогенных РНС. Поскольку нейронная активность РНС формируется в зависимости от весовых коэффициентов и структуры сети в целом, наличие функционально различных модулей может как упростить, так и усложнить декодирование, т.к. влияет на то, как в РНС хранится информация о полученных стимулах. В данной части работы проверялось, есть ли различия при декодировании гомогенных и гетерогенных модификаций РНС.

Использовались по 1500 РНС всех перечисленных ранее гомогенных и гетерогенных модификаций. После обучения задаче реагирования на ряды в течение 1200 ходов записывались доли правильных ответов РНС в режиме теста. Полученный результат соотносится с представленным в разделе 2 (рис. 2А): наибольшая доля правильных ответов в режиме теста в среднем наблюдается у гетерогенных РНС типа RefNet. Точность отклика прочих гетерогенных и гомогенных модификаций РНС имеет близкие значения в диапазоне (0,7-0,8), а модификации РНС с дополнительным входом (SRN+, DTRNN+, RefNet+) в среднем показывают результат чуть хуже, чем без него (SRN, DTRNN, RefNet).

Для дальнейшего исследования было выбрано по 5 РНС каждой модификации, продемонстрировавших наибольшую точность отклика в режиме теста. Результаты декодирования приведены на рис. 3.

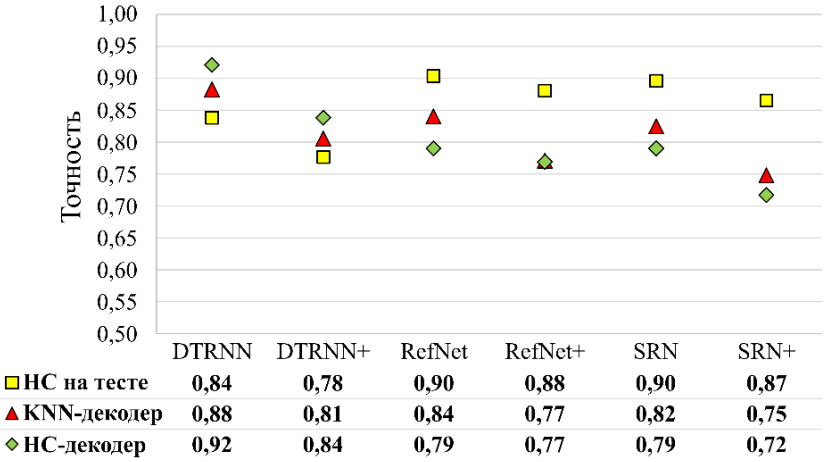

Рис. 3. Средняя точность отклика РНС (доля правильных ответов) в режиме теста, средние точности декодирования нейронной активности этих же РНС с помощью НС- и KNN-декодеров

Было сделано предположение, что декодирование, как процесс извлечения информации паттернов нейронной активности РНС, может сопровождаться потерями. РНС благодаря рекуррентной структуре способны сохранять сведения о контексте (предыдущих тактах), что позволяет им различать даже достаточно похожие между собой фрагменты разных рядов стимулов. Декодеры, в свою очередь, «видят» только картину возбуждения нейронов РНС на текущем такте, т.е. контекст им недоступен. Кроме того, при повторной подаче ряда на вход РНС паттерн нейронной активности может воспроизводиться неточно из-за различий в предыстории, поскольку ряды подаются в случайном порядке, что также может затруднять декодирование. Поэтому ожидалось, что точности декодирования нейронной активности РНС будут ниже, чем точности отклика этих РНС при реагировании на ряды.

На рис. 3 можно видеть, что для РНС типа RefNet, RefNet+, SRN, SRN+ точности декодирования действительно ниже, чем соответствующие точности отклика. Однако для РНС типа DTRNN и DTRNN+ наблюдается обратное: декодирование было проведено с большей точностью, чем само функционирование РНС. Иными словами, данные РНС формируют правильные, хорошо распознаваемые (до 0,92) репрезентации рядов, однако это не всегда приводит к формированию правильного отклика. Точности декодирования нейронной активности быстрого и медленного модулей по отдельности приведены в таблице 3.

Таблица 3. Декодирование отдельных модулей РНС типа DTRNN и DTRNN+

Модуль	Точность KNN-декодирования	
	DTRNN+	DTRNN
РНС целиком	0,889 ± 0,011	0,894 ± 0,002
Быстрый	0,863 ± 0,010	0,883 ± 0,004
Медленный	0,535 ± 0,038	0,563 ± 0,052

Паттерн нейронной активности медленного модуля распознается значительно хуже, чем быстрого. Однако наиболее точное декодирование возможно при рассмотрении всей РНС целиком, что свидетельствует о распределении информации о ряде между быстрым и медленным модулями.

Следовательно, гетерогенность по времени, реализованная в РНС типа DTRNN и DTRNN+, отражается на формировании репрезентаций временных рядов, в то время как функциональная гетерогенность РНС типа RefNet и RefNet+ не приводит к принципиальным отличиям от гомогенных конфигураций. Известно, что поддержание репрезентаций в рабочей памяти реализуется с помощью динамических паттернов [10, 11], что показывает значимость

кодирования во времени. Представленные в настоящей работе результаты показывают, что данное утверждение справедливо и для простейших нейросетевых модельных объектов.

3.3. Оценка сложности паттернов нейронной активности. Декодирование также позволяет оценить сложность паттерна нейронной активности. *Было сделано предположение,* что паттерны нейронной активности гетерогенных РНС обладают большей сложностью по сравнению с гомогенными РНС, поскольку информация распределена между модулями. Оценка сложности возможна, по крайней мере, двумя способами.

Во-первых, в результате предварительных измерений определено, что для обучения НС-декодеров при рассмотрении нейронной активности РНС разных модификаций оптимально разное количество нейронов внутреннего слоя (см. таблицу 4). Наибольший размер внутреннего слоя потребовался для обучения декодированию нейронной активности гетерогенных РНС типа RefNet и RefNet+ (30 нейронов), для прочих РНС значения близки (15-20 нейронов).

Таблица 4. Оптимальное количество нейронов внутреннего слоя НС-декодера

Тип РНС	DTRNN+	DTRNN	RefNet+	RefNet	SRN+	SRN
Количество нейронов внутреннего слоя декодера	20	20	30	30	15	20

Во-вторых, после того, как НС-декодеры обучены, возможно провести редукцию структуры и убрать связи между нейронами (синапсы), не влияющие на качество декодирования [5]. Если проверить, какая часть входных синапсов НС-декодера может быть удалена без потери качества декодирования, то можно оценить, все ли значения нейронной активности РНС на текущем такте функционирования необходимы для идентификации обрабатываемого ряда. Результаты для разных типов РНС приведены на рис.4.

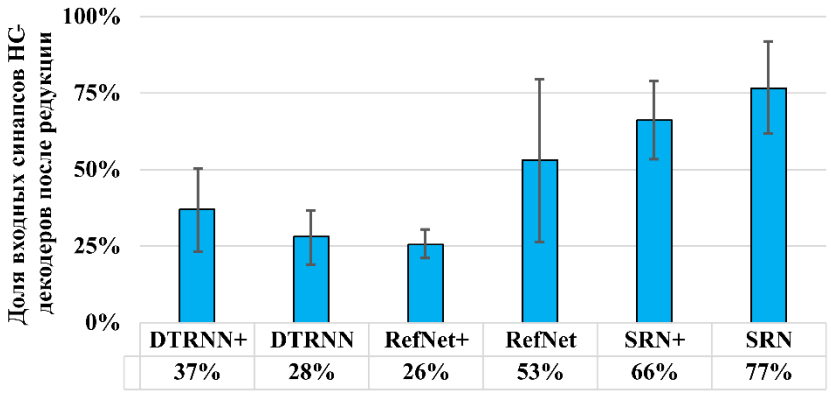

Рис. 4. Доля входных синапсов НС-декодеров, оставшихся после редукции структуры декодеров с сохранением качества декодирования. N=15 для каждой конфигурации, в качестве погрешностей – стандартное отклонение

После редукции НС-декодеры, идентифицировавшие ряды по нейронной активности гомогенных РНС, сохранили в среднем больше входных синапсов по сравнению с гетерогенными. В целом, наименьшую долю входных синапсов (30-40%) сохранили декодеры нейронной активности РНС типа DTRNN и DTRNN+, что согласуется с результатом, приведенным в разделе 3.2: данные РНС обладают наиболее распознаваемыми паттернами нейронной активности. НС-декодеры нейронной активности РНС типа RefNet и RefNet+ также продемонстрировали возможность редукции 50-75% входных синапсов.

Наконец, важно отметить, что паттерны нейронной активности даже в пределах РНС одного типа очень вариативны по сложности, о чем можно судить по величине стандартного отклонения на рис.4. Так, НС-декодер одной сети типа RefNet может редуцироваться на 70-80%, а другой – ни на один синапс.

Заключение. В статье показаны примеры контринтуитивных эффектов, наблюдаемых при моделировании рефлексии на простейших рекуррентных нейронных сетях. Из всех гипотез, приведенных в каждом разделе статьи и связанных с эффективностью и паттернами нейронной активности данных модельных объектов, ни одна не подтвердилась полностью, несмотря на малые размеры и «прозрачное» устройство объектов, а также простоту решаемых ими задач.

Описанные эффекты могут рассматриваться в одном ряду, например, с отображением Фейгенбаума – дискретной формой уравнения Ферхюльста, которое, являясь простым нелинейным уравнением вида $f(x_{n+1}) = rx_n(1 - x_n)$, в зависимости от величины параметра r может описывать разнообразные динамические режимы, в том числе хаотический из-за возникновения каскада бифуркаций [12]. Точки бифуркации, в свою очередь, представляют особый интерес, поскольку, согласно предположению В.А. Лефевра [13, с.10-18], в них поведение одушевленных (обладающих сознанием) тел физически недетерминировано, и наличие таких точек как раз может являться свидетельством наличия сознания. Использованные в работе рекуррентные нейронные сети являются модельными объектами, чье поведение детерминировано уравнениями функционирования (1), однако отклик РНС интерпретируется как «0» или «1» в зависимости от того, на каком из двух выходных нейронов сигнал больше, причем различие между сигналами на этих нейронах может быть сколь угодно малым. Такое различие и может являться источником «*бесконечно малых толчков, направляющих эволюцию состояния тела*» [13, с. 15].

Изложенный здесь материал может служить предостережением для тех исследователей, кто, как и авторы данной статьи, использует предельно простые модельные объекты для воспроизведения реальных феноменов и ожидает предельно ясных результатов, а также для всех использующих в исследовательской практике искусственные нейронные сети.

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

Список источников

1. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science, Behavioral and brain sciences, 2013, vol.36, no.3, pp. 181-204, DOI:10.1017/S0140525X12000477.
2. Beren M., Anil S., Buckley Ch.L. Predictive coding: a theoretical and experimental review. arXiv, 2021, p. 2107.12979, DOI:10.48550/arXiv.2107.12979.
3. Анохин П.К. Избранные труды: философские аспекты теории функциональной системы / П.К. Анохин. – М.: Наука, 1978. – 400 с.
4. Bridges Alice D. et al. Bumblebees socially learn behaviour too complex to innovate alone. Nature, 2024, v.627, no.8004, pp. 572-578, DOI:10.1038/s41586-024-07126-4.
5. Барцев С.И. Нейросетевое декодирование информации о внешнем стимуле по паттерну нейронной активности рекуррентной нейронной сети / С.И. Барцев, П.М. Батурина, Г.М. Маркова // Доклады Российской академии наук. Науки о жизни, 2022. – Т. 502. – № 1. – С. 48-53. – DOI:10.31857/S2686738922010048.
6. Барцев С.И. Биофизический подход к моделированию рефлексии: обоснование, методы, результаты / С.И. Барцев, Г.М. Маркова, А.И. Матвеева // Философские проблемы информационных технологий и киберпространства, 2023. – №.2. – С.120-139. – DOI:10.17726/philIT.2023.2.9.
7. Eliaz K., Rubinstein A. Edgar Allan Poe's riddle: Framing effects in repeated matching pennies games, Games and Economic Behavior, 2011, v.71, no.1, pp. 88-99, DOI:10.1016/j.geb.2009.05.010.
8. Markova G.M., Bartsev S.I. Does a recurrent neural network form recognizable representations of a fixed event series? Advances in Neural Computation, Machine Learning, and Cognitive Research VII. NEUROINFORMATICS 2023, Studies in Computational Intelligence, 2023, vol.1120, pp. 206-213, DOI:10.1007/978-3-031-44865-2_23.
9. Маркова Г.М. Кодирование внешних стимулов простыми рекуррентными нейронными сетями в ходе отложенного теста сравнения / Г.М. Маркова, С.И. Барцев // Сборник научных трудов XXIII Международной научно-технической конференции «Нейроинформатика-2021», Москва, 18-22 октября 2021 года. – Москва: НИЯУ МИФИ, 2021. – С. 39-48. – EDN:ZTPMOY.

10. Stroud J.P., Duncan J., Lengyel M. The computational foundations of dynamic coding in working memory. Trends in cognitive sciences, 2024, vol. 28, iss. 7, pp. 614-627, DOI:10.1016/j.tics.2024.02.011.
11. Cueva Christopher J. et al. Low-dimensional dynamics for working memory and time decoding. PNAS, 2020, vol. 117, no. 37, pp. 23021-23032, DOI:10.1073/pnas.1915984117.
12. May R.M. Simple mathematical models with very complicated dynamics. Nature,1976, vol. 261, no.5560, pp. 459-467, DOI: 10.1038/261459a0
13. Лефевр В.А. Что такое одушевленность? / В.А. Лефевр. – М.: Когито-Центр, 2017. – 123 с.

Маркова Галия Муратовна. *Младший научный сотрудник Института биофизики СО РАН, ассистент, аспирант кафедры биофизики Института фундаментальной биологии и биотехнологии СФУ. Основные направления исследований: нейросетевое моделирование когнитивных функций, нейроинформатика. AuthorID: 1075515, SPIN: 1664-7436, ORCID: 0000-0003-1726-8102, GMarkova@ibp.ru, 660036, Красноярск, ул. Академгородок, 50, стр. 50.*

Барцев Сергей Игоревич. *Доктор физико-математических наук, главный научный сотрудник Института биофизики СО РАН, заведующий лабораторией теоретической биофизики Института биофизики СО РАН, профессор кафедры биофизики Института фундаментальной биологии и биотехнологии СФУ. Основные направления исследований: нейросетевое моделирование когнитивных функций, нейроинформатика, малоразмерные биосферные модели, снижение сложности моделей биологических систем, проектирование биологических систем жизнеобеспечения для космического применения. AuthorID: 66068, SPIN: 1884-5876, ORCID: 0000-0003-0140-4894, BartsevSI@ibp.ru, 660036, Красноярск, ул. Академгородок, 50, стр. 50.*

UDC 577.38+004.81
DOI:10.25729/ESI.2025.37.1.001

Extremely simple does not mean extremely clear: some counterintuitive results of neural network modeling of reflection

Galiya M. Markova[1,2], Sergey I. Bartsev[1,2]

[1]Institute of biophysics Siberian Branch of RAS,
Russia, Krasnoyarsk, *GMarkova@ibp.ru*
[2]School of fundamental biology and biotechnology, Siberian federal university,
Russia, Krasnoyarsk

Abstract. The paper presents results on modeling reflection, understood in a broad sense as the presence of an internal representation of the external world in an active agent that influences its behavior. The ability of the simplest neural network model objects of homogeneous and heterogeneous (modular) structure to solve tasks requiring the presence and use of stable internal representations of external stimuli is revealed. It is determined that these representations are decodable, i.e. based on the current type of neural activity pattern of a neural network, it is possible to determine which specific stimulus or time series of stimuli is currently being processed in it. The authors' initial assumptions made on the basis of general considerations regarding the effectiveness of neural networks of various structures in reflection tasks and the corresponding results are presented. In particular, the following effects are shown: 1) positions in the odd-even game are asymmetric under the condition of limited computational capabilities of the players; 2) formally similar tasks on reflection (the odd-even game and responding to fixed time series of stimuli according to the rules of this game) differ in the requirements for players; 3) decodable patterns of neural activity present not only in neural networks trained to respond to stimuli, but also in networks with random weight coefficients; 4) the accuracy of decoding the neural activity of recurrent neural networks with temporal heterogeneity exceeds the accuracy of the response of these networks when processing series of stimuli; 5) patterns of neural activity in homogeneous recurrent neural networks are more difficult to decode than in heterogeneous networks of comparable size. These effects illustrate the rich internal and behavioral dynamics of the simplest recurrent neural networks, which, on the one hand, is promising for research and practical purposes, and on the other hand, complicates the prediction and interpretation of their behavior.

Keywords: recurrent neural networks, reflection, reflexive games, neural activity decoding, representation of external stimuli

Acknowledgements: The work was supported by a grant from the Russian Science Foundation, Krasnoyarsk Regional Science Foundation No. 23-21-10041 «Hierarchy of functional attractors in neural network models of reflection».

References

1. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science, Behavioral and brain sciences, 2013, vol.36, no.3, pp. 181-204, DOI:10.1017/S0140525X12000477.
2. Beren M., Anil S., Buckley Ch.L. Predictive coding: a theoretical and experimental review. arXiv, 2021, p. 2107.12979, DOI:10.48550/arXiv.2107.12979.
3. Anokhin P.K. Izbrannye trudy: filosofskie aspekty teorii funktsional'noi sistemy [Selected works: philosophical aspects of functional system theory]. Mocsow, Nauka [Science], 1978, 400 p.
4. Bridges Alice D. et al. Bumblebees socially learn behaviour too complex to innovate alone. Nature, 2024, v.627, no.8004, pp. 572-578, DOI:10.1038/s41586-024-07126-4.
5. Bartsev S.I., Baturina P.M., Markova G.M. Neyrosetevoye dekodirovaniye informatsii o vneshnem stimule po patternu neyronnoy aktivnosti rekurrentnoy neyronnoy seti [Neural network-based decoding input stimulus data based on recurrent neural network neural activity pattern]. Doklady Biological Sciences, 2022, vol. 502, no. 1, pp. 1-5, DOI:10.31857/S2686738922010048.
6. Bartsev S.I., Markova G.M., Matveeva A.I. Biofizicheskiy podkhod k modelirovaniyu refleksii: obosnovaniye, metody, rezul'taty [Biophysical approach to modeling reflection: basis, methods, results]. Filosofskiye problemy informatsionnykh tekhnologiy i kiberprostranstva [Philosophical problems of information technology and cyberspace], 2023, no.2, pp. 120-139, DOI:10.17726/philIT.2023.2.9.
7. Eliaz K., Rubinstein A. Edgar Allan Poe's riddle: Framing effects in repeated matching pennies games, Games and Economic Behavior, 2011, v.71, no.1, pp. 88-99, DOI:10.1016/j.geb.2009.05.010.
8. Markova G.M., Bartsev S.I. Does a recurrent neural network form recognizable representations of a fixed event series? Advances in Neural Computation, Machine Learning, and Cognitive Research VII. NEUROINFORMATICS 2023, Studies in Computational Intelligence, 2023, vol.1120, pp. 206-213, DOI:10.1007/978-3-031-44865-2_23.
9. Markova G.M., Bartsev S.I. Kodirovanie vneshnih stimulov prostymi rekurrentnymi neironnymi setjami v hode otlozhennogo testa sravnenija [Encoding of external stimuli by simple recurrent neural networks during a delayed match-to-sample test]. Sbornik nauchnyh trudov XXIII Mezhdunarodnoj nauchno-tehnicheskoj konferentsii «Neuroinformatika-2021», Moskva, 18-22 oktyabrya 2021 goda [Proc. of the XXIII International Scientific and Technical Conference "Neuroinformatics-2021", Moscow, October 18-22, 2021], Moscow, 2021, pp. 39-48, EDN:ZTPMOY..
10. Stroud J.P., Duncan J., Lengyel M. The computational foundations of dynamic coding in working memory. Trends in cognitive sciences, 2024, vol. 28, iss. 7, pp. 614-627, DOI:10.1016/j.tics.2024.02.011.
11. Cueva Christopher J. et al. Low-dimensional dynamics for working memory and time decoding. PNAS, 2020, vol. 117, no. 37, pp. 23021-23032, DOI:10.1073/pnas.1915984117.
12. May R.M. Simple mathematical models with very complicated dynamics. Nature,1976, vol. 261, no.5560, pp. 459-467, DOI: 10.1038/261459a0
13. Lefebvre V.A. Chto takoe odushevlennost'? [What is animateness?]. Moscow, Kogito-Centr, 2017, 123 p.

Markova Galiya Muratovna. *Junior researcher at the Institute of Biophysics SB RAS, assistant, postgraduate student at the Department of biophysics at the Institute of fundamental biology and biotechnology SibFU. Main areas of research: neural network modeling of cognitive functions, neuroinformatics. AuthorID: 1075515, SPIN: 1664-7436, ORCID: 0000-0003-1726-8102, GMarkova@ibp.ru, 660036, Krasnoyarsk, Akademgorodok str., 50, building 50.*

Bartsev Sergey Igorevich. *Doctor of Physical and Mathematical Sciences, Chief Researcher at the Institute of Biophysics SB RAS, Head of the laboratory of theoretical biophysics at the institute of biophysics SB RAS, Professor at the department of biophysics at the Institute of fundamental biology and biotechnology SibFU. Main research areas: neural network modeling of cognitive functions, neuroinformatics, small-scale biosphere models, reducing the complexity of biological systems models, designing biological life support systems for space applications. AuthorID: 66068, SPIN: 1884-5876, ORCID: 0000-0003-0140-4894, BartsevSI@ibp.ru, 660036, Krasnoyarsk, Akademgorodok str., 50, building 50.*

Статья поступила в редакцию 20.12.2024; одобрена после рецензирования 14.01.2025; принята к публикации 13.02.2025.

The article was submitted 12/20/2024; approved after reviewing 01/14/2025; accepted for publication 02/13/2025.