ЖУРНАЛ ВЫСШЕЙ НЕРВНОЙ ДЕЯТЕЛЬНОСТИ, 2023, том 73, № 1, с. 102–123

ДИСКУССИОННЫЕ СТАТЬИ

УДК 159.9.01

СМЕНА ПАРАДИГМЫ В КОГНИТИВНЫХ НАУКАХ?

© 2023 г. Г. Г. Князев*

Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия

**e-mail: knyazevgg@neuronm.ru*

Поступила в редакцию 16.03.2022 г.
После доработки 26.04.2022 г.
Принята к публикации 26.04.2022 г.

Начиная с 50-х годов прошлого века доминирующей парадигмой в когнитивных науках был когнитивизм, который возник как альтернатива бихевиоризму и преимущественно рассматривает когнитивные процессы как разного рода “вычисления” наподобие тех, которые выполняет компьютер, принципиально не отличающийся от универсальной машины Тьюринга (МТ). Несмотря на значительные успехи, достигнутые в последней четверти 20-го века в рамках этой парадигмы, она многих не удовлетворяет, так как не может адекватно объяснить некоторые особенности когнитивных процессов. Возникший позднее коннекционизм, хотя и признает роль вычислительных процессов, их основой считает не МТ, а нейронную сеть, которая лучше моделирует работу мозга, чем вычисления по типу МТ. Нейронные сети, в отличие от классического компьютера, демонстрируют устойчивость и гибкость перед лицом проблем, возникающих в реальном мире, таких как увеличение шума на входе или блокировка части сети. Они также хорошо приспособлены для задач, требующих параллельного разрешения множества противоречивых ограничений. Несмотря на это, аналогия между функционированием человеческого мозга и искусственных нейронных сетей все-таки ограничена в силу радикальных различий в конструкции и связанных с этим возможностей системы. Параллельно с парадигмами когнитивизма и коннекционизма развивались представления, согласно которым когниции являются следствием сугубо биологических процессов взаимодействия организма с внешней средой. Эти представления, которые в последние годы становятся все более популярными, оформились в разные течения так называемого энактивизма. В этом обзоре проводится сравнение теоретических постулатов когнитивизма, коннекционизма и энактивизма, а также парадигмы предсказывающего кодирования и принципа свободной энергии.

Ключевые слова: когнитивизм, коннекционизм, энактивизм, воплощенное познание, предсказывающее кодирование

DOI: 10.31857/S0044467723010094, **EDN:** GJJVJO

ВВЕДЕНИЕ

В настоящее время можно выделить как минимум три конфликтующие парадигмы о природе когнитивных процессов: когнитивизм, коннекционизм и посткогнитивизм. Когнитивизм оформился в 50-х годах прошлого века как альтернатива бихевиоризму и, в противоположность последнему, подчеркивал важность изучения когнитивных процессов. По мере развития когнитивизм оформился в направление, которое преимущественно рассматривает когнитивные процессы как разного рода “вычисления” наподобие тех, которые выполняет компьютер. Соответственно мозг, как орган мышления, рассматривается как вариант сверхсложного компьютера, по устройству в принципе аналогичного современным компьютерам. Неудовлетворенность такой точкой зрения привела к возникновению альтернативной “коннекционистской” парадигмы, согласно которой мозг похож на искусственные нейронные сети. Принципиально другой подход к интерпретации природы когнитивных процессов развивается в ряде современных, достаточно разнородных направлений, которые объединяют общим термином “посткогнитивизм”. Общим в этих направлениях является то, что они, как правило, отвергают аналогии с компьютером и подчеркивают специфически биологическую

основу когнитивных процессов. Несмотря на растущую популярность этих новых направлений, разные варианты когнитивизма и коннекционизма не утратили своего влияния, поэтому я начну обзор с теорий когнитивизма и коннекционизма, а затем рассмотрю наиболее влиятельные направления посткогнитивизма, включая теории предсказывающего кодирования, Байесовской парадигмы, принципа свободной энергии и разных направлений так называемого энактивизма и воплощенного познания.

Когнитивизм

Идеи когнитивизма в наиболее полном виде оформились в направлении, которое обычно описывают термином "вычислительная теория разума" (ВТР, the computational theory of mind). Это направление возникло и развивалось под большим влиянием компьютерных наук и исследований в области искусственного интеллекта (ИИ). Следствием успехов в этой области явилось представление о том, что уже в ближайшем будущем будут созданы компьютеры, ничем не уступающие и даже существенно превосходящие возможности человеческого разума, и что разум вообще – это лишь способность производить вычисления, используя ограниченный набор алгоритмов. Эта идея казалась привлекательной в первую очередь потому, что вычисления, производимые компьютером с использованием алгоритмов, имеют хорошо разработанную математическую базу, которая могла бы позволить описать работу человеческого разума. Прежде чем обсудить основные постулаты ВТР, нужно кратко остановиться на математической теории алгоритмических вычислений.

Алгоритмы, машина Тьюринга и теоремы Геделя

Алгоритм (само слово происходит от имени арабского математика Muḥammad ibn Mūsā al-Khwārizmī, жившего в 9-м веке нашей эры) – это конечная последовательность однозначно интерпретируемых инструкций для решения класса математических проблем или выполнения компьютерных операций (Definition of ALGORITHM. Merriam-Webster Online Dictionary, https://web.archive.org/web/20200214074446/https://www.merriam-webster.com/dictionary/algorithm). Если вся интеллектуальная деятельность человека сводится к использованию алгоритмов, то она не требует какой-либо "сообразительности" или способности "интуитивно" выбрать вариант из множества возможных, она требует лишь знания соответствующих инструкций и их скрупулезного выполнения. Эта идея казалась многим соблазнительной, так как она значительно упрощала понимание процесса мышления, устраняя из него такие туманные явления, как "инсайт" или интуиция. В качестве прототипа интеллектуальной деятельности часто рассматривают математическое мышление, поскольку оно наиболее формализовано и в каком-то смысле может считаться квинтэссенцией и образцом мышления вообще. Во всяком случае, наиболее блистательные достижения научной мысли связаны именно с использованием математики в разных сферах науки.

В 1928 году известный немецкий математик Дэвид Гилберт в соавторстве с Вильгельмом Аккерманом сформулировали так называемую "проблему принятия решений" (по-немецки – Entscheidungsproblem) (Hilbert, Ackermann, 1950): можно ли, используя алгоритмы, решить любую математическую задачу, то есть можно ли для любого адекватно сформулированного математического утверждения решить, верно оно или нет, исходя из набора аксиом и используя правила математической логики? Сам Гилберт был уверен в положительном ответе на этот вопрос.

Дальнейшее развитие теория алгоритмических решений получила в работах Алана Тьюринга, которые имели колоссальное значение для всего последующего развития компьютерных наук, ИИ и ВТР. Принципиально важную роль сыграла предложенная Тьюрингом концепция универсальной вычислительной машины. Машина Тьюринга (МТ), описанная в уже упомянутой выше статье (Turing, 1936), – это абстрактная модель идеализированного вычислительного устройства, имеющего в своем распоряжении неограниченные ресурсы времени и памяти. Машина манипулирует символами, природу которых Тьюринг не конкретизирует. Предполагается, что имеется ограниченный набор простых символов, которые машина может писать или стирать в ячейках памяти. Бесконечное количество ячеек памяти (это могут быть клеточки на бумажной ленте или силиконовые чипы) организовано в линейную последовательность. Центральный процессор в

каждый момент времени имеет доступ лишь к одной ячейке, наподобие сканера, который проходит по бумажной ленте, последовательно сканируя каждую ячейку. Сам процессор может быть в одном состоянии из конечного набора состояний. Процессор может совершать одно из четырех действий: записывать символ в ячейку, стирать символ из ячейки, передвигаться к следующей ячейке ("двигаться по ленте направо"), возвращаться к предыдущей ячейке ("двигаться по ленте налево"). Какое из этих действий будет совершать процессор в каждый момент времени, зависит исключительно от текущего состояния машины и от символа, записанного в актуальной ячейке. Определяющие работу машины правила записаны в таблицу, где указано, какие действия выполняет процессор в зависимости от текущего состояния (из конечного набора состояний) и символа (из конечного набора символов). В таблице также указано, как меняется состояние машины исходя из этих же факторов. Таким образом, в таблице дается ограниченный набор рутинных механических инструкций, управляющих вычислениями. Тьюринг представляет это описание в виде строгой математической модели и утверждает, что такая машина, несмотря на ее простоту, способна выполнить любую выполняемую людьми операцию над конфигурациями символов.

Далее Тьюринг доказывает, что может существовать универсальная МТ (УМТ), которая может воспроизводить поведение любой МТ, если в качестве входной информации она получает таблицу, описывающую соответствующее поведение. УМТ, таким образом, является программируемым компьютером общего назначения. В некотором приближении все современные компьютеры являются УМТ, так как могут выполнять операции любой МТ при наличии соответствующей программы, с той оговоркой, что у них, в отличие от МТ, ограниченный ресурс памяти и времени.

Прогресс компьютерных технологий подталкивал многих, включая Тьюринга, к мысли о том, что в недалеком будущем можно будет создать компьютер, способный мыслить. Это означало бы, что мышление человека ничем принципиально не отличается от алгоритмических операций, выполняемых компьютером. Тьюринг, в частности, предложил способ ответить на вопрос, отличается ли мышление человека от "мышления" компьютера. Он предложил заменить вопрос "может ли компьютер мыслить", который он считал безнадежно туманным, вопросом, может ли компьютер успешно пройти специфический, предложенный им тест. Этот тест, получивший название теста Тьюринга, состоит в том, что эксперт задает вопросы двум невидимым собеседникам, один из которых человек, а другой компьютер. Предлагается считать, что компьютер прошел тест, если эксперт не сможет различить, кто есть кто (Turing, 1950). Философы, однако, критиковали предложение Тьюринга (Block, 1981; Searle, 1980). В частности, Джон Серл выдвинул в качестве контраргумента модель "китайской комнаты" – человек, не знающий китайского языка, получает таблички с текстами на китайском и должен направлять респонденту ответы на этом же языке, следуя подробным инструкциям (написанным на его родном языке). Если инструкции достаточно хороши, у респондента может создаться впечатление, что он общается с человеком, владеющим китайским языком, хотя в реальности его собеседник не понимает ни слова по-китайски (Searle, 1980). Этот аргумент имеет целью показать, что компьютерная имитация, какой бы искусной она ни была, останется лишь имитацией.

Действительно ли человеческое мышление сводится лишь к выведению аксиом и использованию правил логики? Уже в 1931 году австрийский логик Курт Гедель опубликовал две теоремы (Gödel's incompleteness theorems), которые доказывали ограниченность любой аксиоматической системы в ее способности моделировать даже базовые арифметические операции (Gödel, 1931). В теоремах Геделя рассматриваются формальные системы, которые должны соответствовать требованиям полноты и последовательности. Формальная система – это система конечного количества аксиом и правил логики (алгоритмов), позволяющих выводить из аксиом новые теоремы. Такая система считается полной, если любое утверждение, сформулированное с использованием аксиом в соответствии с правилами, или отрицание этого утверждения могут быть выведены в рамках системы. Система считается последовательной (непротиворечивой), если нельзя одновременно вывести и утверждение, и его отрицание. Первая теорема Геделя утверждает, что любая последовательная система, в которой можно производить элементарные арифметические операции, не может быть пол-

ной, то есть в ней есть утверждения, которые нельзя ни доказать, ни опровергнуть. Вторая теорема утверждает, что непротиворечивость любой системы не может быть доказана в рамках самой системы. Этот негативный ответ на сформулированную Гилбертом проблему принятия решений был в дальнейшем независимо подтвержден Алонсо Чёчем (Church, 1936) и Аланом Тьюрингом (Turing, 1936). Теоретический физик, лауреат Нобелевской премии Пенроуз считает, что человеческое мышление имеет в своей основе неалгоритмическую природу, и это позволяет находить решения в условиях неопределенности для прежде неизвестных задач (Penrose, 1989).

Действительно, первоначальный успех в сфере компьютерного моделирования мышления (например, знаменитая программа Logic Theorist, решившая 38 теорем из 52, описанных в Principia Mathematica) сменился периодом застоя, охладившего энтузиазм конструкторов ИИ. Стало понятно, что даже когнитивные процессы относительно низкого уровня, такие как сенсорное восприятие, требуют операций, недоступных современным компьютерам. Одна из трудных проблем, с которой столкнулись исследователи в области ИИ, – это решение задач в условиях неопределенности. В реальной человеческой жизни принятие практически любого решения происходит в условиях неопределенности. В 80-х и 90-х гг. технологический и концептуальный прогресс позволил в какой-то степени преодолеть эти трудности, однако это произошло благодаря использованию новых теоретических и конструкционных идей, в частности, Байесовской теории принятия решений, в которой неопределенность кодируется в терминах вероятности (Murphy, 2012), а также нейронных сетей и машинного обучения. Эти идеи принципиально отличаются от идей классической ВТР и связаны с использованием парадигмы коннекционизма, о которой мы поговорим позже.

Классическая ВТР-репрезентативная теория разума

Классическая ВТР (КВТР) берет свое начало в работе Уоррена Маккаллока и Вальтера Питтса, которые предположили, что что-то, напоминающее машину Тьюринга, может быть хорошей моделью для человеческого разума (McCulloch, Pitts, 1943). В 60-х годах эта модель стала центральной в когнитивной науке, которая исследовала разум методами психологии, ИИ, лингвистики, философии, экономики (особенно теории игр), антропологии и нейробиологии. Согласно КВТР, человеческий разум – это вычислительная машина, принципиально похожая на МТ, и когнитивные процессы, такие как рассуждение и принятие решений, – это вычисления, похожие на вычисления в МТ. В отличие от господствовавших до 60-х гг. теорий, таких как бихевиоризм, который пытался соотнести ментальные состояния с поведенческими паттернами, или теория идентичности, согласно которой ментальные состояния идентичны нейрофизиологическим состояниям, КВТР использует функциональный подход, согласно которому ментальные состояния – это состояния, соответствующие определенной функциональной организации, независимо от субстрата, на котором эта организация реализована (биологическая ткань или, например, чипы компьютера).

Путнам (Putnam, 1967) предложил версию функционализма, которую стали называть машинным функционализмом. По этой версии отдельные когнитивные состояния – это состояния центрального процессора. Таблица состояний определяет функциональную организацию системы и ту роль, которую отдельные состояния в этой организации играют. В отличие от МТ, переходы от состояния к состоянию не детерминированы, а имеют вероятностный характер, то есть разум представляет собой вероятностный автомат. Критикуя теорию Путнама, Блок и Фодор отмечали, что человеческий разум имеет потенциально бесконечное количество состояний, в отличие от конечного количества состояний вероятностного автомата, определяемого таблицей состояний (Block, Fodor, 1972). В предложенной Фодором версии КВТР, которую он назвал “репрезентативная теория разума”, особое внимание уделяется символам, которыми манипулирует разум в процессе Тьюринг-подобных вычислений (Fodor, 1975). Набор этих символов он называет “языком мысли”. Смысл сложных композиций является функцией смысла отдельных частей и способа их комбинирования, который задается набором алгоритмов. КВТР, в частности теория Фодора, нейтральна в отношении субстрата, производящего вычисления. Она потенциально приложима и к дуалистическим интерпретациям, когда вычисления производит Картезианская душа, и к физикалист-

ским интерпретациям, которые субстратом для вычислений видят мозг. На практике все адепты КВТР являются физикалистами и считают, что символы “языка мысли” закодированы в активности нейронов, а алгоритмические операции над символами представляют собой какие-то нервные процессы, природа которых пока неизвестна. Фодор и многие другие приверженцы КВТР считают, что все ментальные процессы сводятся к Тьюринг-подобным вычислениям с использованием набора символов и правил манипулирования этими символами (Gallistel, King, 2009; Fodor, 2005).

Ранние и влиятельные приложения вычислительного подхода к познанию включают теории овладения языком (Chomsky, 1959), внимания (Broadbent, 1958), решения проблем (Newell et al., 1958), памяти (Sternberg, 1969) и восприятия (Marr, 1982). Общей для всех вычислительно-ориентированных исследований является идея о том, что познание включает в себя поэтапную серию событий, начиная с преобразования энергии стимула в символическое выражение, за которым следуют преобразования этого выражения в соответствии с различными правилами, результатом которых является определенный выход – грамматическое языковое высказывание, выделение одного потока слов из другого, решение логической задачи, идентификация стимула как одного из множества запомненных стимулов или трехмерное восприятие мира (Shapiro, Spaulding, 2021).

Символические выражения, над которыми совершают операции когнитивные процессы, а также правила, по которым эти операции выполняются, появляются как репрезентативные состояния познающего агента. Они индивидуализированы в терминах того, о чем они говорят (фонемы, интенсивность света, края, формы и т.д.). Вся эта когнитивная деятельность происходит в нервной системе агента. Именно благодаря активации нервной системы стимулы кодируются в “язык мысли”, подобный языкам программирования, используемым в обычных компьютерах; точно так же правила, диктующие манипуляции с символами в языке мысли, похожи на инструкции, которые выполняет компьютер в процессе выполнения задачи. Методы вычислительной когнитивной науки отражают приверженность этим представлениям. Эксперименты имеют целью выявление содержания репрезентативных состояний или раскрытие шагов, с помощью которых ментальные алгоритмы преобразуют входные данные в выходные.

Несмотря на значительные успехи в понимании когнитивных процессов, достигнутые в последней четверти 20-го века благодаря появлению КВТР, отождествление этих процессов с Тьюринг-подобными вычислениями многих не удовлетворяло. Возникшая в 80-х гг. коннекционистская парадигма, хотя и признает роль вычислительных процессов, ставит под сомнение значение постулируемых КВТР символистских репрезентаций. В следующей главе мы разберем основные идеи, лежащие в основе коннекционизма, который, хотя и противопоставляется классической ВТР, берет начало из тех же источников и в некоторых отношениях на нее похож.

Коннекционизм

Альтернативная КВТР коннекционистская теория (КонТ) появилась в 80-х гг., но, так же как и КВТР, она берет начало в классической работе Уоррена Маккаллока и Вальтера Питтса, которые исследовали сети логических операторов, таких как “И” и “ИЛИ” (McCulloch, Pitts, 1943). Эти сети можно рассматривать в качестве прототипа нейронных сетей, в которых уровень активации узла имеет лишь два значения (0 или 1), а функция активации определяется верностью соответствующего выражения. Маккаллок и Питтс считали логические операторы идеализированной моделью нейронов. Современные цифровые компьютеры фактически представляют собой сети, построенные из логических операторов.

Однако создатели КонТ, в отличие от приверженцев КВТР, черпают вдохновение не в компьютерных науках, а в нейрофизиологии. Основой для вычислений они считают не машину Тьюринга, а нейронную сеть, которая существенно отличается от МТ. Создатели искусственных нейронных сетей использовали в качестве прототипа строение мозга, в частности, устройство зрительного анализатора. Искусственная нейронная сеть – это коллекция связанных друг с другом узлов, которые можно разделить на три категории: входные, выходные и скрытые, расположенные между входными и выходными. Каждый узел в каждый момент времени можно охарактеризовать уровнем активации, выраженным реальным числом, и взвешенными свя-

зями с другими узлами, веса которых (то есть силу связи) можно также выразить реальным числом. Активация входных узлов задается внешним стимулом и служит сигналом для вычислений. Общая активация любого скрытого или выходного узла является функцией суммы взвешенных (на силу связи) активаций питающих его узлов и его собственной активации. Форма этой функции может быть разной для разных сетей. Вычисление в сети представляет собой волну активации, распространяющуюся от входных к выходным узлам в соответствии с весами связей в сети. В прямой сети волна распространяется лишь в одном направлении. В рекуррентных сетях есть петли обратной связи. Рекуррентные сети труднее для математического представления, чем прямые, однако они очень важны для моделирования многих психологических феноменов (Elman, 1990). Веса в нейронной сети обычно могут меняться в соответствии с алгоритмом “обучения”. Описано множество разных алгоритмов, но основная идея состоит в том, чтобы постепенно подстраивать веса так, чтобы результат вычислений в сети приближался к целевому результату, который можно ожидать для соответствующей входной информации. Алгоритм обратного распространения (backpropagation algorithm) является широко используемым примером такого алгоритма (Rumelhart et al., 1986). Таким образом, коннекционистские системы предлагают средства вычислений, которые во многих случаях отказываются от символистских репрезентаций, играющих ключевую роль в КВТР. В отличие от компьютера, который оперирует символами на основе правил, коннекционистские сети преобразуют входные значения активации в выходные значения без использования символов или явных правил, по которым центральный процессор выполняет операции над этими символами.

Сильные стороны коннекционизма

Начиная со второй половины 80-х годов нейронные сети стали привлекать все большее внимание для объяснения разнообразных когнитивных феноменов, включая распознавание объектов, восприятие речи, понимание сентенций и так далее (Rumelhart et al., 1986b). Сторонники КонТ обычно подчеркивают, что нейронные сети гораздо лучше моделируют работу мозга, чем вычисления по типу МТ, поэтому нейронные сети могут обеспечить новую основу для понимания природы разума и его связи с мозгом. Мозг действительно представляет собой нейронную сеть, сформированную из огромного количества единиц (нейронов) и их соединений (синапсов). Более того, некоторые свойства искусственных нейронных сетей позволяют предположить, что коннекционизм может дать гораздо более верное представление о природе когнитивных процессов, чем КВТР.

Устойчивость и гибкость. Нейронные сети демонстрируют устойчивость и гибкость перед лицом проблем, возникающих в реальном мире. Увеличение шума на входе или разрушение части сети приводят к снижению функции, но реакция сети остается адекватной, хотя и несколько менее точной. В отличие от этого в классических компьютерах аналогичные проблемы обычно приводят к катастрофическому отказу.

Параллельное разрешение противоречивых ограничений. Нейронные сети также особенно хорошо приспособлены для решения задач, требующих параллельного разрешения множества противоречивых ограничений. Исследования в области искусственного интеллекта убедительно показывают, что такие когнитивные задачи, как распознавание объектов, планирование и даже координированное движение, представляют собой проблемы такого рода. Модели нейронных сетей обеспечивают гораздо более естественные механизмы для решения таких проблем, чем классические системы (Buckner, Garson, 2019).

Категоризация. На протяжении веков философы пытались понять, как определяются семантические категории. Широко признано, что попытка охарактеризовать их с помощью необходимых и достаточных условий обречена на провал – всегда можно найти исключения практически из любого предложенного определения. Например, можно предположить, что тигр – это большая черно-оранжевая кошка. Но как тогда быть с тиграми-альбиносами? Философы и когнитивные психологи утверждают, что категории разграничиваются более гибкими способами, например, с помощью понятия семейного сходства или сходства с прототипом. Коннекционистские модели кажутся особенно хорошо подходящими для того, чтобы вместить градуированные представления о принадлежности к категориям такого рода. Сети могут научиться оценивать тонкие статистические закономерности, которые очень трудно вы-

разить в виде жестких правил. Коннекционизм обещает объяснить гибкость и проницательность человеческого интеллекта, используя методы, которые не могут быть легко выражены в виде принципов, не допускающих исключений, избегая тем самым ненадежности стандартных форм символического представления (Horgan, Tienson, 1990).

Кодирование смысла в мозге. Коннекционистские модели представляют собой новую парадигму для понимания того, как информация может быть представлена в мозге. Какое-то время назад соблазнительная, но наивная идея заключалась в том, что отдельные нейроны (или небольшие компактные группы нейронов) могут кодировать каждый объект, или концепцию (так называемая "клетка бабушки", grandmother cell, которая отвечает только на образ бабушки) (Barlow, 1994). Однако эмпирические данные показывают, что любая мысль сопровождается активностью большого количества нейронов, широко распределенных по разным участкам коры головного мозга (см. обзор Князев, 2022). Интересно отметить, что распределенные, а не локальные представления на скрытых единицах нейронных сетей являются естественным продуктом коннекционистских методов обучения (Buckner, Garson, 2019). Распределенные представления (в отличие от символов, хранящихся в отдельных фиксированных местах памяти) относительно хорошо сохраняются при разрушении или перегрузке части сети. Что еще более важно, поскольку представления кодируются в виде паттернов, а не активности отдельных единиц, отношения между представлениями кодируются в сходствах и различиях между этими паттернами. Таким образом, внутренние свойства репрезентации несут информацию о том, что она кодирует (Clark, 1993). В отличие от этого локальная репрезентация является условной. Никакие ее внутренние свойства не определяют ее содержание. В схеме символического представления все представления состоят из символических атомов (как слова в языке). Значения сложных строк символов могут определяться тем, как они построены из своих составляющих, но что фиксирует значения слов? Коннекционистские схемы репрезентации смысла позволяют обойти эту загадку, просто отказавшись от атомов. Каждое распределенное представление – это паттерн активности всех единиц, поэтому нет принципиального способа отличить простые представления от сложных. Конечно, репрезентации складываются из активности отдельных единиц. Но ни один из этих "атомов" не кодирует какой-либо символ – репрезентации не являются символическими. Более того, наличие самих "репрезентаций" и "языка мысли", постулируемых КВТР, являются предметом дебатов (Buckner, Garson, 2019). В ряде работ Хорган и Тиенсон (Horgan, Tienson, 1990) отстаивали точку зрения, называемую репрезентациями без правил. Согласно этой точке зрения, классики правы, считая, что человеческий мозг (и его хорошие коннекционистские модели) содержит репрезентации, но они ошибаются, считая, что эти репрезентации подчиняются жестким правилам, как шаги компьютерной программы. Идея о том, что коннекционистские системы могут следовать градуированным или приблизительным закономерностям ("мягким законам", как их называют Хорган и Тиенсон), интуитивно понятна и привлекательна.

Достоинства и недостатки распределенных репрезентаций. Одна из привлекательных черт распределенных репрезентаций в коннекционистских моделях заключается в том, что они предлагают решение проблемы кодирования смысла в активности мозга: сходства и различия между паттернами активации "кодируют" семантическую информацию. Таким образом, сходства нейронных активаций обеспечивают свойства, которые определяют смысл (Buckner, Garson, 2019). Однако развитие теории смысла, основанной на сходстве паттернов активации, сталкивается с серьезными препятствиями (Fodor, Lepore, 1999), поскольку такая теория должна была бы объяснять смысл предложения на основе анализа смысла отдельных его частей, и неясно, как одно лишь сходство паттернов активации способно решать такие задачи в том виде, как этого требует стандартная теория. Тем не менее большинство коннекционистов, которые продвигают основанные на сходстве объяснения смысла, отвергают многие из предпосылок стандартных теорий. Они надеются создать рабочую альтернативу, которая либо отвергает, либо изменяет эти предпосылки, оставаясь при этом способной объяснить данные о лингвистических способностях человека. Кальво (Calvo, 2003) считает, что в этом коннекционисты потерпят неудачу, так как не смогут решить проблему сопутствующей информации. Эта проблема заключается в том, что сходство между паттернами ак-

тивации для какого-то понятия (скажем, “бабушка”) в двух человеческих мозгах наверняка будет очень низким, потому что у двух людей (сопутствующая) информация о бабушках (имя, внешность, возраст, характер) будет очень разной. Если понятия определяются всем, что мы знаем, то паттерны активации этих понятий будут очень далеки друг от друга. Это, однако, большая проблема для любой теории, которая надеется определить смысл концепций через функциональные состояния мозга, будь то традиционная или коннекционистская парадигма.

Слабые стороны коннекционизма

Несмотря на эти интригующие особенности, в коннекционистских моделях есть некоторые недостатки, о которых стоит упомянуть.

Отличие нейронных сетей от мозга. Реальное устройство мозга гораздо сложнее, чем конструкция любой из современных нейронных сетей. Большинство исследований нейронных сетей абстрагируются от многих интересных и, возможно, важных особенностей мозга. Нейроны гораздо более гетерогенны, чем узлы сети, в частности, они имеют разную нейрохимическую природу. Кроме того, они сочетают аналоговый мембранный потенциал с дискретными потенциалами действия. Самое важное отличие состоит в том, что большинство алгоритмов машинного обучения (например, алгоритм обратного распространения) требуют наличия эталона, то есть результата, к которому нужно стремиться в процессе обучения. Этот эталон программа получает от оператора, знающего правильный ответ. Ничего подобного нет в работе мозга. В некоторых нейронных сетях вместо алгоритма обратного распространения используется алгоритм обучения с подкреплением (Pozzi et al., 2019) и другие алгоритмы, для которых не нужен эталон (Krotov, Hopfield, 2019), но они пока менее успешны. Кроме того, коннекционистские модели пока не способны предложить правдоподобные механизмы долговременного хранения информации в мозге, кроме петель реверберирующей активности, которые безнадежно неэффективны. Далеко не очевидно также, что мозг содержит такие обратные связи, которые были бы необходимы, если бы мозг обучался с помощью такого процесса, как обратное распространение, а огромное количество повторений, необходимое для таких методов обучения, кажется нереалистичным в приложении к мозгу. Внимание к этим вопросам, вероятно, необходимо, если мы хотим построить убедительные коннекционистские модели человеческих когнитивных процессов.

Моделирование высших когнитивных функций. Более серьезное возражение состоит в том, что, хотя нейронные сети неплохо справляются с моделированием когнитивных процессов нижнего уровня, таких как распознавание и категоризация сенсорной информации, они не особенно хороши для моделирования процессов, которые лежат в основе языка, рассуждений и высших форм мышления (Pinker, Prince, 1988). Критики коннекционизма утверждают, что коннекционистские модели хороши только для обработки ассоциаций, но когнитивные способности высокого уровня, такие как владение языком и рассуждение, не могут быть реализованы с использованием только ассоциативных методов, поэтому коннекционисты вряд ли смогут сравниться с классическими моделями в объяснении этих способностей. Фодор и Пилишин (Fodor, Pylyshyn, 1988) выделяют особенность человеческого интеллекта, называемую систематичностью, которую, по их мнению, коннекционисты не могут объяснить. Систематичность языка относится к тому факту, что способность производить/понимать/думать одни предложения неразрывно связана со способностью производить/понимать/думать другие, имеющие сходную структуру. Например, ни один человек, понимающий фразу “Иван любит Марию”, не может не понимать фразу “Мария любит Ивана”. С классической точки зрения это можно легко объяснить, если предположить, что знание языка предполагает знание смысла слов (“Иван”, “любить” и “Мария”) и грамматику фразы “Иван любит Марию”, и это знание позволяет “вычислить” значение фразы из значений этих составляющих. Если это так, то понимание нового предложения “Мария любит Ивана” может быть объяснено как еще один случай того же символического процесса. Аналогичным образом, символическая обработка объясняет систематичность рассуждений, обучения и мышления. Фодор и Маклафлин (Fodor, McLaughlin, 1990) доказывают, что, хотя коннекционистские модели можно обучить систематичности, их также можно обучить, например, распознавать фразу “Иван любит Марию”, не будучи в состоянии распознать фразу “Мария

любит Ивана”. Поскольку коннекционизм не гарантирует систематичность, он не объясняет, почему систематичность так широко распространена в человеческом познании. Систематичность может существовать в коннекционистских архитектурах, но там, где она есть, это не более чем счастливая случайность. Классическое решение намного лучше, потому что в классических моделях повсеместная систематичность возникает естественно. Дебаты по этому вопросу продолжаются по сей день, но в целом кажется очевидным, что его трудно решить с позиций радикального коннекционизма и более предпочтительны какие-то гибридные варианты (Calvo, Symons, 2014).

Различие стратегий мышления. В 2010-х годах в компьютерных науках особую популярность приобрели модели, известные как “глубокие нейронные сети”, или “сети глубокого обучения”, которые содержат большое количество скрытых узлов (иногда сотни слоев) и тренируются на больших базах данных с использованием того или иного алгоритма обучения (LeCun et al., 2015). Использование этих сетей позволило получить замечательные результаты во многих областях ИИ, таких как распознавание объектов и стратегические игры, и они сейчас широко используются в коммерческих приложениях. Их также используют для моделирования когнитивных процессов (Marblestone et al., 2016). Наиболее значительный прогресс в сфере создания все более эффективных искусственных нейронных сетей связан, в частности, с появлением так называемых глубоких сверточных сетей. Например, программа AlphaZero обыгрывает мировых чемпионов в трех различных играх (шахматы, Сёги и Го) “без знания человеческой стратегии”, то есть используя только информацию о правилах этих игр и стратегии, которые она находит в процессе интенсивной игры сама с собой (Silver et al., 2018). Это достигается с помощью средств, недоступных человеку (например, AlphaZero провела более 100 миллионов партий игры в Го сама с собой), что ставит вопрос о правомерности использования этих сетей в качестве модели человеческого мозга. Необычный подход AlphaZero к стратегии произвел мини-революцию в изучении шахмат и Го (Sadler, Regan, 2019) и вызвал опасения, что решения, которые обнаруживают глубокие сети, для человека являются чуждыми и загадочными. Сложная структура глубоких сверточных сетей затрудняет объяснение их решений в конкретных случаях. Эта проблема породила движение “объяснимого ИИ”, цель которого – вдохновить разработку инструментов для анализа решений компьютерных алгоритмов, особенно для того, чтобы системы ИИ могли быть сертифицированы на соответствие практическим или юридическим требованиям (Goodman, Flaxman, 2017). Необходимость в объяснимых глубоких сетях становится еще более актуальной в связи с обнаружением так называемых “конфронтационных примеров” (Nguyen et al., 2015). Они бывают как минимум в двух формах: “пертурбированные изображения”, которые представляют собой естественные фотографии, очень незначительно измененные таким образом, что вызывают резкие изменения в классификации глубоких сетей, хотя разница незаметна для человека, и “мусорные изображения”, которые бессмысленны для человека, но классифицируются глубокими сетями с высоким уровнем уверенности. Эти примеры привели некоторых к выводу, что если у сети есть “понимание” объектов, которыми она манипулирует, то это понимание должно радикально отличаться от человеческого. Все это показывает, что вряд ли стоит идти слишком далеко в аналогиях между устройством и функционированием искусственных нейронных сетей и человеческого мозга.

Различия и сходство коннекционизма и КВТР

Таким образом, если согласно КВТР человеческое познание аналогично символическим вычислениям в цифровых компьютерах (информация представлена строками символов, точно так же, как мы представляем данные в памяти компьютера или на листе бумаги), то согласно коннекционизму информация хранится не символически, в весах, или силе связи, между единицами нейронной сети. Классицист считает, что познание похоже на цифровую обработку, где строки создаются последовательно в соответствии с инструкциями символической программы. Коннекционист рассматривает когнитивную обработку как динамическую и градуированную эволюцию активности в нейронной сети, где активация каждой единицы зависит от силы связи и активности ее соседей. На первый взгляд, эти позиции кажутся несовместимыми, однако многие коннекционисты не рассматривают свою работу как вызов классицизму, а некоторые открыто поддер-

живают классическую картину. Они стремятся найти компромисс между двумя парадигмами и считают, что сеть мозга реализует символьный процессор на более высоком и абстрактном уровне описания. Действительно, хотя КВТР и КонТ часто рассматривают как альтернативные теории, они по сути не являются взаимоисключающими – нейронные сети могут моделироваться и классической МТ, как это делается в современных цифровых компьютерах и, с другой стороны, классические вычисления по типу машины Тьюринга могут осуществляться и с использованием нейронных сетей (Graves et al., 2014). Радикальные коннекционисты утверждают, что символическая обработка была с самого начала плохим предположением о том, как работает разум. Они считают, что классическая теория плохо справляется с объяснением гибкости и эффективности человеческого познания, и хотели бы навсегда исключить символическую обработку из когнитивной науки (Buckner, Garson, 2019). Более умеренные коннекционисты предлагают объединить идеи коннекционизма с идеями КВТР, создавая так называемые гибридные коннекционистские архитектуры, включающие в нейронные сети элементы классической символьной обработки (Wermter, Sun, 2000).

Вопрос, на который пока не может ответить ни КВТР, ни КонТ, – это как мозг, построенный из относительно гораздо медленнее (по сравнению с компьютером) работающих элементов (достаточно вспомнить время синаптической задержки и время распространения импульса по аксону), может осуществлять такие сложные вычисления так быстро и эффективно (Gallistel, King, 2009). Наконец, пока нет понимания того, как вычислительные теории разума могут объяснить каузальную силу содержания (семантики) сознания. Очевидно, что традиционные вычислительные системы манипулируют символами на основе синтактических правил, описываемых алгоритмами (например, таблица состояний в машине Тьюринга). Они, таким образом, “слепы” в отношении семантики, то есть смысла символов, наподобие того, как обитатель “китайской комнаты” “слеп” в отношение смысла китайских фраз, которыми он манипулирует. Адепты КВТР настаивают на том, что семантические интерпретации “языка мыслей” не влияют прямо на результаты вычислений, то есть не имеют каузальной силы. Результаты вычислений зависят исключительно от формальных правил манипулирования символами, то есть разум является “синтактической машиной” (Fodor, 1980). Сторонники КонТ более осторожны в отрицании роли семантики, но в большинстве своем также согласны с тем, что вычисления имеют преимущественно синтаксический характер. Многие философы критикуют этот тезис сторонников КВТР и КонТ и доказывают, что поведение человека и результаты его рассуждений принципиально зависят от семантики его мыслей (Block, 1990; Figdor, 2009; Kazez, 1995). Некоторые критики вычислительных теорий разума доказывают, что возможности человеческого разума превосходят возможности алгоритмических машин (Lucas, 1961; Nagel, Newman, 1958; Penrose, 1989). Интуиция, креативность, инсайт не могут быть смоделированы на основе алгоритмов, так же как и многие другие особенности человеческого познания (Dreyfus, 1992). Даже такие апологеты КВТР, как Фодор, часто выражают (особенно в его последних трудах) скептицизм в отношении того, что КВТР может объяснить все важные свойства человеческого разума (Fodor, 2000).

В целом, вычислительные теории разума представляют интересную попытку формального описания когнитивных операций. Они являются своего рода интерфейсом между когнитивными и компьютерными науками и пока, похоже, более полезны для последних, чем для первых. Появившиеся в процессе развития этих теорий и их приложений наработки, такие как различные варианты нейронных сетей и машинного обучения, получили широкое применение в разных сферах и позволили достичь существенного прогресса в области ИИ. Эти достижения сейчас находят применение в том числе и при изучении и моделировании активности мозга и когнитивных процессов. Успехи, достигнутые при использовании нейронных сетей глубокого обучения в решении таких задач, как классификация объектов по семантическим категориям и распознавание зрительных образов, показывают, что похожие механизмы может использовать и мозг для решения аналогичных задач. Сомнительно, однако, что КВТР, КонТ и их варианты смогут описать и объяснить все особенности человеческого сознания и разума.

“Обработка информации” в мозге

При рассмотрении вычислительных теорий разума нельзя обойти вниманием стандартное среди когнитивных психологов описание активности мозга термином “обработка информации” (information-processing). Понятие информации стало очень популярным начиная с работ Клайда Шеннона, заложившего в 1948 г. основы математической теории коммуникации (Shannon, 1948). Шеннон определял информацию как меру уменьшения неопределенности, которая выражается в виде изменения распределения вероятностей возможных состояний (Cover, Thomas, 2006). Есть и другие определения информации. Фред Дретске определяет информацию, содержащуюся в какой-либо переменной, через ее корреляцию с другой переменной (Dretske, 1981). Например, количество колец на срезе дерева коррелирует с его возрастом, значит, оно содержит информацию о возрасте дерева. Еще одно определение – “семантическая информация” – относится к содержанию репрезентации (например, репрезентация объекта внешнего мира в сознании человека) (Fodor, 1998; Sprevak, 2010). Понятие информации приобрело колоссальное значение в разнообразных науках, включая физику (например, квантовая теория информации), многие разделы инженерных наук, компьютерные науки, ИИ и когнитивные науки (например, теория интегрированной информации). Ряд ученых и философов, начиная, пожалуй, с Джона Уиллера (Wheeler, 1989), предлагают считать информацию одной из основ мироздания, наряду с материей (Chalmers, 1996; Davis, 2010; Gleick, 2011; Lloyd, 2007; Seife, 2007; Vedral, 2010). При этом значение, в котором используется термин “информация”, часто не оговаривается. Можно отметить, что информация, в каком бы смысле она ни определялась, имеет значение лишь для существа, способного ее извлечь и понять. Легко увидеть, что для обладающего сознанием и разумом человека информация в разных ее видах имеет первостепенное значение. Можно думать, что собака, обнюхивающая забор со следами, оставленными другими собаками, извлекает информацию, имеющую для нее значение. Менее очевидно, но потенциально возможно использование понятия информации для объяснения поведения растений или бактерий. В отношении неорганической природы, если речь не идет о ее описании человеком, понятие информации теряет смысл. Поэтому философ Джон Серл считает объяснение феномена сознания с помощью информации примером циркулярного мышления, поскольку само понятие информации предполагает наличие обладающего сознанием существа, способного ее понять (Searle, 2013).

Предсказывающее кодирование

Далее я кратко опишу основные направления и представления, развиваемые различными вариантами посткогнитивизма, но перед этим необходимо упомянуть об очень влиятельной и амбициозной парадигме “предсказывающего кодирования” (predictive coding). Я не буду здесь подробно останавливаться на этой большой теме. В общих чертах, согласно теории предсказывающего кодирования, восприятие сенсорной информации является активным процессом. Мозг постоянно генерирует и обновляет ментальную модель окружающего мира. На основании этой модели генерируется предсказание (ожидание) сенсорного входного сигнала, которое сравнивается с реальным сигналом, и рассогласование между ожиданием и реальностью (ошибка предсказания) используется для коррекции ментальной модели. Идея предсказывающего кодирования при восприятии внешнего мира высказана Гельмгольцем еще в 1860 году, разрабатывалась американским психологом Джеромом Брюнером в 1940-х гг. и стала предметом экспериментальных исследований после опубликования в 1981 г. статьи МакКлеланда и Румелхарта (McClelland, Rumelhart, 1981).

На интуитивном уровне идею предсказывающего кодирования легко понять из повседневного опыта. Если, например, человек входит в хорошо знакомое помещение, он ожидает увидеть хорошо ему знакомые предметы на привычных местах и, если картина соответствует ожиданию, он ее практически не замечает (нулевая ошибка рассогласования). Если же вопреки ожиданию он обнаружит посреди комнаты новый и неизвестный ему предмет (большая ошибка рассогласования), все внимание будет направлено на то, чтобы понять, что это за предмет и объяснить его появление в этом месте. Современные модели предсказывающего кодирования постулируют механизмы, основанные на Байесовском подходе и воплощенные в мозге в виде иерархических многослойных сетей с

восходящими и нисходящими связями между слоями (Clark, 2013).

Байесовская модель

Байесовские вероятностные модели принятия решений в условиях неопределенности приобретают все большую популярность в когнитивных науках. Можно кратко рассмотреть этот подход на примере сенсорного восприятия, которое лучше других когнитивных процессов соответствует Байесовской модели. Сенсорное восприятие – классический образец принятия решений в условиях неопределенности. Еще Гельмгольц подчеркивал, что проблема неопределенности является эндемической для сенсорного восприятия (Helmholtz, 1867). Например, у зрительного анализатора нет непосредственного доступа к отдаленному окружению, у него есть доступ только к сенсорным стимулам, возникающим в сетчатке. Как зрительный анализатор принимает решение о свойствах отдаленного объекта на основе ограниченной информации, поставляемой сетчаткой? Гельмгольц предполагал, что стимулы, поступающие из сетчатки, запускают неосознаваемый процесс принятия в качестве решения наиболее правдоподобной гипотезы о свойствах отдаленного объекта. Байесовский подход к моделированию сенсорного восприятия основан на этом предположении Гельмгольца. Байесовская модель восприятия включает пространство гипотез, в котором каждая гипотеза h относится к какому-либо аспекту воспринимаемого объекта (форма, размер, цвет и так далее). Исходная вероятность $p(h)$ – это начальное правдоподобие, приписываемое h. Исходная вероятность сенсорного сигнала e в свете гипотезы h оценивается как $p(e|h)$. После получения сигнала e система пересчитывает вероятность h с учетом e. По теореме Байеса апостериорная вероятность рассчитывается как $p(h|e) = \eta p(h)p(e|h)$, где η – это размерностная константа, необходимая для того, чтобы все вероятности суммировались в единицу. На основе апостериорной вероятности система выбирает наиболее правдоподобную гипотезу. Этот выбор управляется функцией максимизации полезности, которая оценивает цену ошибки (Rescorla, 2019).

Сейчас “Байесовскую когнитивную науку” выделяют даже в виде отдельной дисциплины, которая рассматривает Байесовские модели восприятия, моторного контроля, причинных рассуждений, социального познания, интуиции, навигации и анализа естественного языка. Предполагается, что в основе всех этих процессов лежит вероятностный анализ информации Байесовского типа, который происходит на подсознательном уровне. Критики Байесовской парадигмы указывают, что большое количество ментальных процессов протекает не по Байесовскому типу. Канеман и Тверский доказывают, что рассуждения на сознательном уровне обычно не используют исчисления вероятностей, а при принятии решений часто не учитывается ожидаемая максимизация пользы (Kahneman, Tversky, 1979; Tversky, Kahneman, 1983). Некоторые считают, что и на подсознательном уровне когнитивные процессы часто не подчиняются законам Байесовской парадигмы (Morales et al., 2015). Кроме того, Байесовская модель не объясняет, откуда берутся гипотезы. Можно думать, что гипотезы генерируются с учетом ситуационного контекста (предсказывающее кодирование) на основе информации, хранящейся в памяти, однако, как рассчитывается их исходная вероятность и как все это закодировано в активности мозга, остается неизвестным (Orlandi, 2016). Сторонники Байесовского подхода соглашаются, что конкретные механизмы пока неясны и что вряд ли можно объяснить все когнитивные процессы на основе теории Байеса, но некоторые, особенно сенсорное восприятие и моторный контроль, лучше всего описываются именно в рамках этой теории (Rescorla, 2019). Одна из самых больших проблем теории заключается в смутном понимании того, как именно минимизация ошибок предсказания может быть устроена в мозге. Ключевой нерешенный вопрос – что именно представляет собой сигнал ошибки и как он рассчитывается на каждом уровне обработки сигнала в мозге (Bastos et al., 2012). В некоторых фМРТ-исследованиях увеличение BOLD-сигнала интерпретируется как сигнал ошибки, в других же – как входной сигнал (Kogo, Trengove, 2015). Другая проблема состоит в том, что вычисления, которые согласно модели должны производиться на каждом уровне иерархической сети, могут быть компьютационно неразрешимыми (Kwisthout, van Rooij, 2019).

Принцип свободной энергии

В последние десятилетия, в основном трудами Карла Фристона и соавторов (Friston et al., 2006), идеи предсказывающего кодирования и Байесовская модель были объединены с некоторыми другими идеями в общую теорию, названную “принцип свободной энергии” (ПСЭ, free energy principle). Эта теория предлагается как “объединяющая теория мозга”, способная интегрировать экспериментальные данные, относящиеся к восприятию, действию и обучению, и даже еще шире – как теория поведения любой биологической системы. ПСЭ постулирует, что любая самоорганизующаяся система, которая находится в равновесии с окружающей средой, должна минимизировать свою свободную энергию. По сути, это математическая формулировка того, как адаптивная система (то есть любая биологическая система) противостоит естественной тенденции к увеличению энтропии (в соответствии со вторым законом термодинамики). Согласно ПСЭ, поведение биологического агента направлено на уменьшение “удивления”, то есть рассогласования между ожидаемой и реальной сенсорной стимуляцией. Это рассогласование, которое и называется свободной энергией, является функцией сенсорных состояний и так называемой “плотности узнавания” (вероятностная репрезентация возможных причин, вызвавших сенсорную стимуляцию). Агенты могут уменьшать свободную энергию, изменяя две вещи, от которых она зависит: они могут изменять сенсорную информацию, воздействуя на окружающий мир, или они могут изменить свою плотность узнавания, изменив внутреннюю модель мира. Математически в основе ПСЭ лежит Байесовская модель мозга, описывающая восприятие как конструктивный процесс, опирающийся на внутреннюю генеративную модель мира, которую мозг старается оптимизировать, используя входные сенсорные сигналы.

Посткогнитивизм

Несмотря на теоретическую плодовитость и впечатляющие успехи на ранних этапах своего существования, когнитивистская парадигма столкнулась с некоторыми трудноразрешимыми проблемами. Кажущиеся интуитивно наиболее простыми когнитивные способности, такие как контроль моторики и восприятие, оказались наименее податливыми для объяснения в рамках этой парадигмы. Кроме того, среди философов отсутствовал консенсус по поводу понятия “репрезентация”, лежащего в основе когнитивизма. Неудовлетворенность когнитивизмом была причиной возникновения альтернативных теорий, которые явились предтечей современных направлений посткогнитивизма (смотри Ward et al., 2017).

Предтечи посткогнитивизма

Работы коннекционистов с нейронными сетями показывают, что объяснение разумного поведения не нуждается в обращении к серийному производству и манипулированию дискретными репрезентативными состояниями – адаптивное поведение может возникать из активности сети взаимодействующих единиц. Важно отметить, что паттерны связности, которые определяют структуру сети коннекционистов, необязательно должны быть жесткими или заранее заданными. Вместо этого коннекционистские сети могут быть самоорганизующимися системами; структура, лежащая в основе их разумного поведения, может возникать в результате обучения сети и ее интерактивной истории. Коннекционистские модели имели заметный успех в областях, которые для когнитивистов представляли особую проблему, таких как распознавание образов и сенсомоторный контроль.

Для объяснения организации когнитивных процессов стали привлекать теорию динамических систем (ТДС), которая предоставляет формальный концептуальный аппарат для описания развертывания операций сложных систем, состоящих из нескольких тесно взаимодействующих частей, включая самоорганизующиеся системы, такие как (некоторые) сети коннекционистов (Horgan, Tienson, 1992). Язык ТДС характеризует системы с точки зрения многомерного пространства возможных состояний, в которых система может находиться, уравнений, описывающих способы, которыми система может переходить от одной точки в пространстве состояний к другой, и теоретически значимые точки в этом пространстве, такие как аттракторы – стабильные состояния, к которым система стремится. Эти характеристики не привлекают понятия дискретных, статических репрезентаций в пользу глобального описания состояния системы и ее активности.

Параллельно с этим работы в области экологической психологии развивали альтернативный когнитивизму взгляд на природу сенсорного восприятия (Chemero, 2011). Согласно экологической психологии Гибсона (Gibson, 1979), зрительное восприятие окружающей среды является “прямым”, в том смысле, что его не следует понимать в терминах репрезентативных состояний или вычислительных операций, восстанавливающих информацию об окружающей среде, которая теряется при сенсорной трансдукции. Одной из причин, почему такие состояния не нужны, является активность восприятия. Сенсорное воздействие окружающей среды разворачивается во времени и может модулироваться нашей собственной деятельностью (прищуривание, более внимательное рассматривание, перемещение). Когнитивистская концепция зрительного восприятия как восстановления детальной информации из статического и обедненного восприятия недооценивает ресурсы, доступные нашим сенсорным системам. Во-вторых, то, что мы воспринимаем, связано с нашими целями и возможностями. Мы воспринимаем “аффордансы” – возможности взаимодействовать с окружающей средой таким образом, чтобы это отражало наши потребности и планы, а не практически нейтральную информацию, которую наши системы восприятия должны интерпретировать и уже потом сопоставлять с нашими способностями к действию. Эта концепция восприятия представляет воспринимающего и окружающую среду как созависимые системы. Окружающей средой для воспринимающего субъекта является тот набор свойств окружения, которые могут направлять его текущую деятельность, и быть таким субъектом – значит быть существом, которое может руководствоваться этими свойствами окружения.

С практической точки зрения представления экологической психологии развивали энтузиасты так называемой “локальной” робототехники (situated robotics), пионером которой был Родни Брукс (Brooks, 1991). Брукс отмечал, что построенные на принципах когнитивизма роботы не могут даже приблизиться к воспроизведению адаптивного поведения простых насекомых. Брукс создал серию роботов, которых он назвал “Существа” (Creatures), способных производить набор простых адаптивных действий в процессе взаимодействия с реальным окружением. В отличие от когнитивистского принципа моделирования процесса постоянно обновляемой детальной репрезентации окружающей обстановки, Брукс снабдил свои Существа набором специализированных подсистем, большинство из которых управляли простым сенсомоторным поведением. Вместо того чтобы соединять эти подсистемы с центральным процессором, который должен вычислять единый план действий робота и управлять его поведением в соответствии с этим планом, подсистемы были взаимосвязаны таким образом, чтобы деятельность каждой могла подавлять деятельность других способами, которые инженер мог легко подстраивать под текущие нужды. Существа Брукса предоставили, таким образом, доказательство того, что простые варианты осмысленного поведения можно воспроизвести с помощью нескольких взаимодействующих сенсомоторных модулей, не прибегая к детальной репрезентации окружающей среды и централизованного контроля (замечу в скобках, что для осмысленности поведения тут все-таки потребовалось участие инженера).

В философии истоки когнитивизма совпадают с истоками “аналитической философии”, согласно которой логические построения, лежащие в основе мышления, можно описать в терминах формальных правил манипулирования синтаксическими структурами, и мысль конструируется в виде пропозициональной установки. Работы некоторых феноменологов, однако, предложили альтернативную концепцию мышления. Так, Хайдеггер (Heidegger, 1927/1962) считал, что способность мыслителя представлять элементы своего окружения (как в пропозициональной установке) зависит от предшествующей способности умело взаимодействовать с окружающей средой способами, которые подчиняются нормативным ограничениям. Мерло-Понти (Merleau-Ponty, 1945/2012) аналогичным образом утверждал, что способность поддерживать осмысленные когнитивные отношения с окружающей средой зависит от способности к телесному взаимодействию с ней, и детали этого взаимодействия вносят решающий вклад в структуру мысли и субъективного опыта. Все эти идеи витали в воздухе, когда разрабатывался первый манифест энактивизма. В противовес акценту когнитивизма на конструкции и манипуляции дискретными внутренними репрезентациями, каждое из этих направлений по-разному подчеркивало объяснительную силу взаимо-

действия со средой и важность внешних факторов и факторов реализации поведения, которые могут включать свойства тела познающего, его окружение и динамику взаимодействий между этими факторами.

Энактивизм

Программа энактивизма в целостном виде впервые была представлена в книге Френсиса Варелы с соавторами в 1991 году (Varela et al., 1991). В двух предложениях, авторы так описывают эту программу: “энактивистский подход состоит из двух пунктов: (1) восприятие состоит в действии, управляемом восприятием, и (2) когнитивные структуры возникают из повторяющихся сенсомоторных паттернов, которые позволяют действовать, руководствуясь восприятием” (Varela et al., 1991, стр. 173). Для понимания отличия между когнитивизмом и энактивизмом второй пункт особенно важен. Для когнитивизма когнитивные структуры являются внутренними состояниями, которые представляют определенные свойства окружающей среды. Энактивизм вместо этого делает упор на эмерджентные когнитивные структуры, которые самоорганизуются в результате взаимодействия между организмом и окружающей средой. Ключевым для энактивизма является представление об организме как об “аутопоэтической” системе, которая “генерирует и определяет свою собственную организацию, действуя как система, производящая собственные компоненты” (Maturana, Varela, 1980, стр. 79). Таким образом, единство биологической системы возникает и поддерживается в процессе взаимодействий с окружающей средой. Детали этих взаимодействий (их отличительная динамика) зависят от устройства организма и от свойств окружающей среды, способствующих процветанию организма. Например, хотя сахароза является реальным компонентом физико-химической среды бактерии, ее статус пищи таковым не является. То, что сахароза является питательным веществом, не присуще статусу молекулы сахарозы как таковой; это свойство сахарозы является следствием взаимоотношения бактерии со средой и связано с метаболизмом бактерий. Значение сахарозы как пищи создается самим организмом (Thompson, 2007). Энактивистский подход, таким образом, подразумевает, что и сам организм, и значимые структуры в его окружении возникают из набора самоорганизующихся динамических процессов. Эти структуры окружающей среды значимы лишь постольку, поскольку они влияют на успех или неудачу организма в самосохранении как аутопоэтической целостности, и в этом смысле имеют значение для существования организма. В силу этого структуры, участвующие в таких взаимодействиях, могут считаться когнитивными (Ward et al., 2017). Именно такой взгляд на совместное производство познающего и окружающей среды посредством динамического взаимодействия определяет взгляд на восприятие как действие, управляемое восприятием. Варела и соавторы считают, что “Существа” Брукса дают наглядный пример энактивистской когнитивной науки (Varela et al., 1991) в силу акцента Брукса на взаимодействии с окружающей средой, в противовес построению детальных внутренних репрезентаций. Точно так же очевидно сходство между энактивизмом и экологическим подходом Гибсона, который делает упор на взаимозависимости организма и окружающей среды и на способности организма к прямому взаимодействию со структурами, которые влияют на успех его деятельности.

Акцент энактивизма на вовлеченной деятельности, а не на отстраненном представлении, и способ, которым детали устройства организма определяют детали познавательного отношения к окружающей среде – две точки соприкосновения с описанной выше феноменологической традицией. Другая – это отказ от реалистической и объективистской концепции мира, в пользу концепции мира как продукта и отражения нашей деятельности. Энактивизм провозглашается как современное продолжение феноменологической традиции Мерло–Понти в стремлении найти золотую середину между реалистическими и идеалистическими концепциями взаимоотношений между разумом и миром (Varela et al., 1991). Энактивисты считают, что отстраненное постфактум теоретизирование может исказить переживания, которые оно стремится анализировать. Они ратуют за внедрение в когнитивную науку Буддийской практики медитации и внимательности (mindfulness), чтобы увести разум от его теорий и предвзятостей к непосредственной ситуации самого чувственного опыта. Ключевое утверждение Варела (Varela et al., 1991) состоит в том, что различные когнитивные научные тенденции, синтезированные в энактивизме, поддер-

живают видение ментальности как эмерджентной, воплощенной и вовлеченной сущности, которое было замечено феноменологами, такими как Мерло–Понти, и занимает центральное место во многих буддийских традициях. Различные варианты энактивизма включают аутопоэтический энактивизм (Thompson, 2007), сенсомоторный энактивизм (O'Regan, 2011), радикальный энактивизм (Hutto, Myin, 2012) и различные попытки понять ментальное как нечто воплощенное (embodied, вовлекающее не только мозг, а все телесные структуры и процессы), встроенное (embedded, функционирующее только в соответствующей внешней среде), действенное (enacted, вовлекающее не только нервные процессы, но и то, что организм делает), расширенное (extended, в окружающую среду организма) и аффективное (affective) (Rowlands, 2010). Энактивисты утверждают, что любая система, которая имеет автономию, самореференцию и способность себя создавать и поддерживать, обладает когнитивными способностями. Следовательно, познание присутствует во всех живых системах. Радикальные энактивисты делают упор на отрицании ментальной репрезентации внешнего мира. Они приходят к выводу, что базовые когниции, а также когниции простых организмов, таких как бактерии, лучше всего охарактеризовать как нерепрезентативные. Что касается сложных форм познания у человека, таких как язык, они появились в полной форме только с созданием социокультурных когнитивных ниш в человеческом обществе (Hutto, Myin, 2012).

Одно из возражений энактивистскому подходу к познанию состоит в том, что он не может объяснить более сложные формы когнитивных способностей, такие как человеческие мысли, которые чрезвычайно трудно объяснить без репрезентаций (Clark, Toribio, 1994). Сторонники энактивизма пытаются ответить на это возражение с помощью интеграции концепций энактивизма и предсказывающего кодирования (ПК). Теория ПК, которую мы рассмотрели выше, утверждает, что любое познание регулируется необходимостью свести к минимуму ошибку предсказания. Сами предсказания, как правило, основываются на эволюционных потребностях, так что организмы ожидают получения сенсорной информации, которая помогает им поддерживать гомеостатическую жизнеспособность. Например, люди ожидают, что они окажутся в ситуациях, когда получаемые данные помогают им поддерживать температуру тела, процент воды, уровень кислорода и т.д., необходимые для сохранения жизни (Downey, 2020). Если они окажутся в среде, в которой эти ожидания не оправдываются – например, у жерла извергающегося вулкана – они используют когнитивные способности, чтобы снизить расхождение между ожидаемыми и фактическими сенсорными данными. Например, они могут задерживать дыхание, чтобы не вдыхать вулканический пепел и газ, и стремиться удалиться как можно дальше от вулкана. В двух словах, согласно теории ПК, организмы поддерживают гомеостаз, предпринимая действия, необходимые для снижения отклонений от ожидаемого положения дел. Важным следствием ПК является то, что исчезают традиционные различия между когнитивными категориями, такими как "восприятие", "аффект" и "действие", – все познание направляется стремлением уменьшить ошибки рассогласования, а перцептивные, аффективные и поведенческие стратегии считаются разными способами достижения этого (Downey, 2020). Вместо того чтобы говорить о различных перцептивных, аффективных и поведенческих системах, мы должны говорить об одной общей когнитивной системе, которая позволяет организму обладать ориентированной на действие и аффективно нагруженной перцептивной связью со своим окружением.

Попытаемся суммировать основные области расхождения между когнитивистской и энактивистской парадигмами. В значительной степени отличия обусловлены тем, что используется в качестве парадигмального прототипа когнитивной системы. Когнитивизм рассматривает разум или мозг, с которым он, как правило, идентифицируется, как компьютер, который получает информацию о внешнем мире через сенсорные системы и на основе этой информации строит модели внешнего мира (репрезентации), которые затем уже используются для управления поведением в соответствии с текущими потребностями организма. Важно подчеркнуть, что репрезентации создаются независимо от потребностей организма. Они могут быть ложными или однобокими в силу несовершенства органов чувств и доступности информации, и задачей центрального процессора является их исправление для уменьшения рассогласования с реальной картиной мира, но само создание репрезентаций не ограни-

чивается уже на входе потребностями организма. В этой схеме можно выделить отдельные относительно независимые компоненты. Прежде всего, это организм и внешний мир, который считается объективно существующим и независимым от организма. В организме можно выделить мозг (управляющий орган) и остальное тело, подчиняющееся мозгу и выполняющее функции снабжения мозга информацией и выполнения его команд – для поддержания гомеостаза и осуществления необходимого поведения. В мозге, соответственно, тоже есть блоки переработки сенсорной информации, регуляции движения и вегетативной регуляции и центральный процессор, осуществляющий координацию на основе создаваемых им репрезентаций внешнего мира. Это, естественно, грубая и утрированная схема, но она отражает общий подход когнитивизма.

Отправной пункт энактивизма – это биологический организм, являющийся аутопоэтической системой. Парадигмальным прототипом когнитивной системы в энактивизме является бактерия. Различие между биологическим организмом и компьютером очевидно. У компьютера нет собственных эгоистических интересов – он создан человеком для выполнения его заданий. Соответственно, создаваемая им "модель мира" в общем-то объективна (в меру доступности информации, несовершенства воспринимающих систем и ошибок центрального процессора). Аутопоэтическая система, в отличие от компьютера, сама себя создает и поддерживает. Поскольку система эта находится в неравновесном состоянии, она должна прикладывать постоянные усилия для его сохранения и противостояния физическим силам, спонтанно направленным на ее разрушение. У этой системы поэтому есть свой интерес, и он состоит не в создании объективной модели мира, а в успешном взаимодействии с этим миром, позволяющем сохранить собственную целостность. Предполагается, что сама эта система формировалась в процессе эволюции как продукт взаимодействия с внешним миром (с той его частью, которая является экологической нишей для данного организма). В этом смысле "модель" внешнего мира запечатлена в самом строении организма – "организм является воплощенной теорией своего окружения" (Munz, 2002). Эта модель мира, выраженная в анатомической структуре организма и его инстинктивных поведенческих паттернах, накладывает жесткие ограничения на вид, объем и качество получаемой через сенсорные каналы информации. Например, информация о мире радикально различна у дождевого червя, не имеющего зрения, кошки, способной видеть в темноте, летучей мыши, обладающей эхолокацией, и человека. Но это не все. Согласно энактивизму, организм не получает пассивно всю информацию, доступную его органам чувств, а активно выбирает лишь то, что необходимо для поддержания его жизнедеятельности (аффордансы). Сам процесс получения информации неразрывно связан с поведенческим взаимодействием с миром. Собственно, термин "информация" в представлении энактивистов по сути лишен смысла в приложении к сенсорному восприятию. Этот процесс жизнедеятельности невозможно разбить на компоненты – традиционное различие между восприятием, аффектом и действием исчезает и заменяется единой когнитивной системой, обеспечивающей ориентированную на действие и эмоционально нагруженную перцептивную "хватку". В качестве метафоры такой когнитивной системы используют образ спагетти Эшера, представляющих собой нити, концы которых возвращаются в их собственные начала, превращая "вход" и "выход", а также "ранний" и "поздний" в неточные и вводящие в заблуждение понятия (Clark, 2009).

В этом представлении понятие репрезентации теряет смысл, и отрицание репрезентации как когнитивного состояния является основным посылом радикального энактивизма. По сути, теряет смысл понятие внешнего мира вообще, так как организм имеет дело лишь с аффордансами. Что собой представляет "реальный" независимый от организма мир – остается за пределами рассмотрения. Мы, как биологические организмы, имеем дело лишь с результатами нашего взаимодействия с избранными в соответствии с нашими возможностями и потребностями аспектами этого мира. В этом смысле энактивизм напоминает прагматику Копенгагенской интерпретации квантовой механики. Для радикального энактивизма отрицание репрезентации является принципиальным в его борьбе с когнитивизмом, что особенно заметно в том, как трактуются концепция предсказывающего кодирования (ПК) и принцип свободной энергии. Эти концепции постулируют наличие в мозге моделей, таких как ге-

неративная Байесовская модель, на основе которых мозг делает предсказания предстоящего сенсорного входа. Эти модели апостериорно усовершенствуются для уменьшения рассогласования между предсказанием и реальным сенсорным входом. Почему мы не можем считать эти модели репрезентацией, с той поправкой, что они моделируют не собственно реальный внешний мир, а сенсорное отражение этого мира? Апологеты энактивизма настаивают, однако, что нет необходимости в наличии внутренней модели для осуществления ПК. Оно может происходить на основе "чувства" рассогласования (Downey, 2020), или оценки корреляции между ожидаемой и реальной сенсорной стимуляцией (Kirchhoff, Robertson, 2018), и динамического корректирования соответствия действий организма его потребностям при взаимодействии с окружающей средой (Gallagher, Bower, 2013). На основе чего, однако, возникает "чувство" рассогласования и делается предсказание в Байесовском ПК – остается неясным.

Биологизм энактивизма, его упор на сенсомоторное взаимодействие с окружением для поддержания собственной жизнедеятельности хорошо подходит для описания поведения бактерии – излюбленного объекта в ранних трудах энактивистов, но труднее прилагается к описанию когнитивных процессов человека (Clark, Toribio, 1994). В более поздних трудах идеи энактивизма распространили на когнитивную активность высшего уровня, присущую человеку, которую энактивисты рассматривают как взаимодействие с социальной средой (Di Paolo et al., 2014). Это распространяется на мысли и знание, в том числе научное знание. "С энактивистской точки зрения ... знание конструируется: оно конструируется агентом посредством его сенсомоторных взаимодействий с окружающей средой, совместно конструируется между и внутри каждого человека посредством их значимого взаимодействия друг с другом. В своей наиболее абстрактной форме знание конструируется между людьми в социолингвистических взаимодействиях ... Наука – это особая форма конструирования социальных знаний ... [которая] позволяет нам воспринимать и предсказывать события за пределами нашего непосредственного познавательного понимания ...а также для построения дальнейших, еще более мощных научных знаний" (Rohde, 2010, стр. 30). В этом взгляде на знание энактивизм сближается с конструктивизмом. Однако конструктивизм рассматривает интерактивность как радикальный, творческий, ревизионистский процесс, в котором знающий конструирует личную "систему знаний" на основе своего опыта и проверяет ее жизнеспособность в практических встречах с окружающей средой. Обучение является результатом воспринимаемых аномалий, вызывающих неудовлетворенность существующими представлениями (Von Glasersfeld, 1989). Можно ли считать, что человеческое знание, в том числе научное знание, формировалось исключительно потребностью самосохранения человека как биологической системы? Безусловно, практический (как говорят, прикладной) аспект неизбежно есть в этом знании, однако есть и то, что называют фундаментальной наукой, непосредственная связь которой с биологическими потребностями человека не так очевидна. В целом, энактивизм представляется прагматически полезным пересмотром традиционного когнитивизма при условии смягчения его наиболее радикальных постулатов.

В настоящее время накоплен уже немалый объем эмпирических данных, в целом согласующихся с постулатами энактивизма, в частности, с его вариантом, получившим название "воплощенная когниция" (embodied cognition). Концепция воплощенной когниции (КВК) постулирует, что смысл символов и концепций имеет корни в нашем опыте взаимодействия с внешним миром, и предсказывает, что доступ к концептуальному знанию должен задействовать те же процессы, которые активны при получении или непосредственном использовании этого знания (Shapiro, 2019). Эти предсказания хорошо согласуются с данными фМРТ-исследований кодирования конкретных концепций в активности мозга. Меньше пока уверенности в том, что КВК сможет адекватно объяснить кодирование абстрактных концепций (для обзора этих работ см. (Князев, 2022)).

ЗАКЛЮЧЕНИЕ

Можно отметить общую тенденцию эволюции представлений о природе ментальных процессов при переходе от когнитивизма к коннекционизму и далее к энактивизму. Когнитивизм рассматривает их в виде формальных операций над символами, которые принципиально не отличаются от операций компьютера на основе жестко прописанных

алгоритмов. В основе своей эта точка зрения является редукционистской: наши ментальные состояния – это состояния центрального процессора, их можно описать последовательностью вычислительных операций, как это прописано в программе, выполняемой компьютером. Реальное устройство мозга и его биологическая основа в расчет не берутся. Да, мы пока не знаем, как конкретно вычислительные операции закодированы в активности мозга, но, по большому счету, это не имеет значения. Такой подход позволяет описать некоторые особенности когнитивных процессов человека, но в целом построенные на его основе модели далеки от реальности.

Коннекционизм пытается отойти от этой схемы и строит модели когнитивных операций на основе формализованных представлений об устройстве мозга. Этот шаг позволяет обойти многие трудности, непреодолимые для КВТР, и создать модели, гораздо более похожие на реальность, хотя некоторые их свойства продолжают казаться чужеродными. Энактивизм идет еще дальше по пути “биологизации” когнитивных процессов, постулируя, что суть когнитивных состояний неразрывно связана с процессом взаимодействия организма с окружающей средой. Отметим, что ни коннекционизм, ни энактивизм не тяготеют к редукционизму. Содержание ментальных состояний человека невозможно вывести, или свести к распределенной активности нейронных сетей, или к сенсомоторным взаимодействиям организма с окружением для поддержания собственной жизнедеятельности. Оно в данном случае рассматривается как эмерджентная сущность.

ФИНАНСИРОВАНИЕ

При написании этой статьи автор получал финансовую поддержку Российского научного фонда (проект № 22-15-00142).

СПИСОК ЛИТЕРАТУРЫ

Князев Г.Г. Кодирование смысла в активности мозга. Журнал высшей нервной деятельности им. И.П.Павлова. 2022. В печати.

Barlow H.B. The neuron doctrine in perception. In M. Gazzaniga (Ed.), The Cognitive Neurosciences. 1994. Boston: MIT Press.

Bastos A.M., Usrey W.M., Adams R.A., Mangun G.R., Fries P., Friston K.J. Canonical microcircuits for predictive coding. Neuron. 2012. 76: 695–711.

Block N. Psychologism and Behaviorism. Philosophical Review. 1981. 90: 5–43.

Block N. Can the Mind Change the World? in Meaning and Method: Essays in Honor of Hilary Putnam. *G. Boolos* (ed.), 1990. Cambridge: Cambridge University Press.

Block N., Fodor J. What Psychological States Are Not. The Philosophical Review. 1972. 81: 159–181.

Broadbent D.E. Perception and Communication. 1958. New York: Pergamon Press.

Brooks R.A. Intelligence without representation. Artificial Intelligence. 1991. 47 (1): 139–159.

Buckner C., Garson J. Connectionism. The Stanford Encyclopedia of Philosophy. 2019. E.N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.

Calvo G.F. Connectionist Semantics and the Collateral Information Challenge. Mind & Language. 2003. 18 (1): 77–94.

Calvo P., Symons J. The Architecture of Cognition: Rethinking Fodor and Pylyshyn’s Systematicity Challenge. 2014. Cambridge: MIT Press.

Chalmers D.J. The Conscious Mind. 1996. Oxford University Press.

Chemero A. Radical Embodied Cognitive Science. 2011. Cambridge, MA: MIT Press.

Chomsky N. On Certain Formal Properties of Grammars. Information and Control. 1959. 2 (2): 137–167.

Church A. An unsolvable problem of elementary number theory. American Journal of Mathematics. 1936. 58: 345–363.

Clark A. Associative Engines: Connectionism, Concepts, and Representational Change. 1993. Cambridge, MA: MIT Press.

Clark A. Spreading the joy? Why the machinery of consciousness is (probably) still in the head. Mind. 2009. 118 (472): 963–993.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences. 2013. 36 (03), 181–204.

Clark A., Toribio J. Doing without representing. Synthese. 1994. 101 (3): 401–434.

Cover T., Thomas J. Elements of Information Theory. 2006. Hoboken: Wiley.

Davis P. Information and the Nature of Reality. 2010. Cambridge, MA: MIT Press.

Di Paolo A.E., Rhohde M., De Jaegher H. Horizons for the enactive mind: Values, social interaction, and play. In J. Stewart, O. Gapenne & E.A Di Paolo (eds.). Enaction: Toward a New Paradigm for Cognitive Science. 2014. Cambridge, MA: MIT Press.

Downey A. It Just Doesn’t Feel Right: OCD and the ‘Scaling Up’ Problem. Phenomenology and the Cognitive Sciences. 2020. 19 (4): 705–727.

Dretske F. Knowledge and the Flow of Information. 1981. Oxford: Blackwell.

Dreyfus H. What Computers Still Can't Do. 1992. Cambridge, MA: MIT Press.

Elman J. Finding Structure in Time. Cognitive Science. 1990. 14: 179–211.

Figdor C. Semantic Externalism and the Mechanics of Thought. Minds and Machines. 2009. 19: 1–24.

Fodor J. The Language of Thought. 1975. New York: Thomas Y. Crowell.

Fodor J. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. Behavioral and Brain Science. 1980. 3: 63–73.

Fodor J. Concepts. 1998. Oxford: Clarendon Press.

Fodor J. The Mind Doesn't Work That Way. 2000. Cambridge, MA: MIT Press.

Fodor J. Reply to Steven Pinker 'So How Does the Mind Work?'. Mind and Language. 2005. 20: 25–32.

Fodor J., Lepore E. Holism: A Shopper's Guide. 1992. Cambridge: Blackwell.

Fodor J., McLaughlin B.P. Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work. Cognition. 1990. 35 (2): 183–204.

Fodor J.A., Pylyshyn Z.W. Connectionism and Cognitive Architecture: A Critical Analysis. Cognition. 1988. 28 (1–2): 3–71.

Friston K., Kilner J., Harrison L. A free energy principle for the brain. Journal of Physiology, Paris. 2006. 100 (1–3): 70–87.

Gallagher S., Bower M. Making enactivism even more embodied. Avant: Trends in Interdisciplinary Studies. 2013. 2.

Gallistel C.R., King A. Memory and the Computational Brain. 2009. Malden: Wiley-Blackwell.

Gibson J.J. The Senses Considered as Perceptual Systems. 1966. Boston: Houghton Mifflin.

Gibson J.J. The Ecological Approach to Visual Perception. 1979. Boston: Houghton Mifflin.

Gleick J. The Information. 2011. Pantheon.

Gödel K. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. Monatshefte für Mathematik und Physik. 1931. 38: 173–198.

Goodman B., Flaxman S. European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'. AI Magazine. 2017. 38 (3): 50–57.

Graves A., Wayne G., Danihelko I. Neural Turing Machines. arXiv preprint. 2014. arXiv:1410.5401.

Heidegger M. Being and time. Translated by J. Macquarrie and E. Robinson. 1927/1962. Basil Blackwell: Oxford.

Helmholtz H. von. Handbuch der Physiologischen Optik. 1867. Leipzig: Voss.

Hilbert D., Ackermann W. Principles of Mathematical Logic. 1950. AMS Chelsea Publishing, Providence: Rhode Island, USA.

Horgan T.E., Tienson J. Soft Laws. Midwest Studies in Philosophy. 1990. 15: 256–279.

Horgan T.E., Tienson J.L. Cognitive systems as dynamic systems. Topoi. 1992. 11 (1): 27–43.

Hutto D., Myin E. Radical Enactivism: Basic Minds Without Content. 2012. Cambridge, MA: MIT Press.

Kahneman D., Tversky A. Prospect theory: An analysis of decision under risk. Econometrica. 1979. 47: 263–291.

Kazez J. Computationalism and the Causal Role of Content. Philosophical Studies. 1995. 75: 231–260.

Kirchhoff M.D., Robertson I. Enactivism and predictive processing: A non-representational view. Philosophical Explorations. 2018. 21(2): 264–281.

Kogo N., Trengove C. Is predictive coding theory articulated enough to be testable? Frontiers in Computational Neuroscience. 2015. 9: 111.

Krotov D., Hopfield J. Unsupervised Learning by Competing Hidden Units. Proceedings of the National Academy of Sciences, USA. 2019. 116: 7723–7731.

Kwisthout J., van Rooij I. Computational Resource Demands of a Predictive Bayesian Brain. Computational Brain & Behavior. 2019. 3 (2): 174–188.

LeCun Y., Bengio Y., Hinton G. Deep Learning. Nature. 2015. 521: 436–444.

Lloyd S. Programming the Universe. 2007. Vintage.

Lucas J.R. Minds, Machines, and Gödel. Philosophy. 1961. 36: 112–137.

Marblestone A., Wayne G., Kording K. Toward an Integration of Deep Learning and Neuroscience. Frontiers in Computational Neuroscience. 2016. 10: 1–41.

Marr D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. 1982. San Francisco: W.H. Freeman.

Maturana H.R., Varela F.J. Autopoiesis and cognition: the realization of the living. 1980. Kluwer: Dordrecht.

McClelland J.L., Rumelhart D.E. An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological Review. 1981. 88 (5): 375–407.

McCulloch W., Pitts W. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics. 1943. 7: 115–133.

Merleau-Ponty M. Phenomenology of perception. Translated by *D.A. Landes*. 1945/2012. Routledge: London.

Morales J., Solovey G., Maniscalco B., Rahnev D., de Lange F., Lau H. Low attention impairs optimal incorporation of prior knowledge in perceptual

decisions. Attention, Perception, and Psychophysics. 2015. 77: 2021–2036.

Munz P. Philosophical Darwinism: On the Origin of Knowledge by Means of Natural Selection. 2002. Routledge: London.

Murphy K. Machine Learning: A Probabilistic Perspective. 2012. Cambridge, MA: MIT Press.

Nagel E., Newman J.R. Gödel's Proof. 1958. New York: New York University Press.

Newell A., Shaw J.C., Simon H.A. Elements of a Theory of Human Problem Solving. Psychological Review. 1958. 65 (3): 151–166.

Nguyen A., Yosinski J., Clune J. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). 2015. 427–436.

O'Regan J.K. Why red doesn't sound like a bell. Explaining the feel of consciousness. 2011. Oxford University Press: Oxford.

Orlandi N. Bayesian perception is ecological perception. Philosophical Topics. 2016. 44: 327–351.

Penrose R. The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics. 1989. Oxford University Press, UK.

Pinker S., Prince A. On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. Cognition. 1988. 28 (1–2): 73–193.

Pozzi I., Bohté S., Roelfsema P. A Biologically Plausible Learning Rule for Deep Learning in the Brain. arXiv preprint. 2019. arXiv:1811.01768.

Putnam H. Psychophysical Predicates. In Art, Mind, and Religion, *W. Capitan and D. Merrill* (eds), Pittsburgh: University of Pittsburgh Press. 1967, 429–440.

Rescorla M.A. Realist Perspective on Bayesian Cognitive Science. In: *T. Chan and A. Nes* (Eds) Inference and Consciousness. 2019. Routledge: London.

Rohde M. The scientist as observing subject. Enaction, Embodiment, Evolutionary Robotics: Simulation Models for a Post-Cognitivist Science of Mind. 2010. Atlantis Press. pp. 30.

Rowlands M. The new science of the mind: from extended mind to embodied phenomenology. 2010. MIT Press: Cambridge.

Rumelhart D., Hinton G., Williams R. Learning Representations by Back-propagating Errors. Nature. 1986. 323: 533–536.

Rumelhart D., McClelland J., and the PDP Research Group. Parallel Distributed Processing. Vol. 1. 1986b. Cambridge: MIT Press.

Sadler M., Regan N. Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI. 2019. Alkmaar: New in Chess.

Searle J. Minds, Brains and Programs. Behavioral and Brain Sciences. 1980. 3: 417–457.

Searle J. Can Information Theory Explain Consciousness? The New York Review. 2013. 10.

Seife C. Decoding the Universe. 2007. Penguin.

Shapiro L. Embodied cognition. 2019. Routledge: London.

Shapiro L., Spaulding S. Embodied Cognition. The Stanford Encyclopedia of Philosophy. 2021. E.N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/>.

Silver D., Hubert T., Schrittwieser J., Antonoglou I., Lai M., Guez A., Lanctot M., Sifre L., Kumaran D., Graepel T., Lillicrap T., Simonyan K., Hassabis D. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play. Science. 2018. 362 (6419): 1140–1144.

Sprevak M. Computation, Individuation, and the Received View on Representation. Studies in History and Philosophy of Science. 2010. 41: 260–270.

Sternberg S. Memory-Scanning: Mental Processes Revealed by Reaction-Time Experiments. American Scientist. 1969. 57 (4): 421–457.

Thompson E. Mind in Life: Biology, Phenomenology, and the Sciences of Mind. 2007. Cambridge, MA: Harvard University Press.

Tversky A., Kahneman D. Extension versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. Psychological Review. 1983. 90: 293–315.

Turing A. On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, Series 2. 1936. 42: 230–265.

Turing A. Computing Machinery and Intelligence. Mind. 1950. 49: 433–460.

Varela F.J., Thompson E., Rosch E. The embodied mind. 1991. MIT Press: Cambridge.

Vedral V. Decoding Reality. 2010. Oxford University Press, UK.

Von Glasersfeld E. Cognition, construction of knowledge and teaching. Synthese. 1989. 80 (1): 121–140.

Ward D., Silverman D., Villalobos M. Introduction: The varieties of enactivism. Topoi. 2017. 36 (3): 365–375.

Wermter S., Sun R. (eds.). Hybrid Neural Systems. (Lecture Notes in Computer Science 1778). 2000, Berlin, Heidelberg: Springer Berlin Heidelberg.

Wheeler J.A. Information, physics, quantum: the search for links. Proceedings III International Symposium on Foundations of Quantum Mechanics, Tokyo. 1989, 354–368.

PARADIGM CHANGE IN COGNITIVE SCIENCES

G. G. Knyazev[#]

Federal State Budgetary Scientific Institution "Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia

[#]*e-mail: knyazevgg@neuronm.ru*

Since the 1950s, the dominant paradigm in the cognitive sciences has been cognitivism, which emerged as an alternative to behaviorism, and predominantly views cognitive processes as various kinds of "computations" similar to those performed by the computer. Despite significant advances made in the last quarter of the 20th century within this paradigm, it does not satisfy many scientists because it could not adequately explain some features of cognitive processes. Connectionism, which emerged somewhat later, recognizes the role of computational processes, but as their basis considers a neural network, which is a much better model of brain functioning than Turing-type computations. Neural networks, unlike the classical computer, demonstrate robustness and flexibility in the face of real-world problems, such as increased input noise, or blocked parts of the network. They are also well suited for tasks requiring the parallel resolution of multiple conflicting constraints. Despite this, the analogy between the functioning of the human brain and artificial neural networks is still limited due to radical differences in system design and associated capabilities. Parallel to the paradigms of cognitivism and connectionism, the notions that cognition is a consequence of purely biological processes of interaction between the organism and the environment have developed. These views, which have become increasingly popular in recent years, have taken shape in various currents of the so-called enactivism. This review compares the theoretical postulates of cognitivism, connectionism, and enactivism, as well as the predictive coding paradigm and the free energy principle.

Keywords: cognitivism, connectionism, enactivism, embodied cognition, predictive coding