ОБЩИЕ ВОПРОСЫ ФОРМАЛИЗАЦИИ ПРОЕКТИРОВАНИЯ: ОНТОЛОГИЧЕСКИЕ АСПЕКТЫ И КОГНИТИВНОЕ МОДЕЛИРОВАНИЕ

УДК 004.5 *Научная статья* DOI: 10.18287/2223-9537-2024-14-4-466-482

О машинном обучении, мифах о сильном искусственном интеллекте и о том, что такое понимание

© 2024, О.П. Кузнецов

Институт проблем управления РАН, Москва, Россия

Аннотация

В первой части статьи обсуждается книга американского учёного Э. Ларсона «Миф об искусственном интеллекте», которая посвящена разоблачению мифов об искусственном интеллекте. Эти мифы, история которых насчитывает более половины века, заключаются в том, что возникновение человекоподобного («сильного») искусственного интеллекта, а в дальнейшем и сверхинтеллекта якобы неизбежно, и оно произойдёт как бы само собой – в результате постепенной эволюции систем искусственного интеллекта. Критика этих мифов проводится в книге по двум направлениям: научному и социальному. Показано, что машинное обучение не ведёт к сильному искусственному интеллекту, а миф об искусственном интеллекте ослабляет веру в человеческий потенциал. Во второй части статьи рассматривается проблема понимания. Предлагается концепция когнитивной семантики, основанная на идеях Дж. Лакоффа, С. Пинкера, А. Дамасио и А. Сета. В частности отмечается, что: понимание – это интерпретация в терминах картины мира человека; картину мира строит наш мозг, и она структурируется через категоризацию опыта человека; значения (смыслы) формируются раньше, чем формируются понятийные структуры; в основе значений лежат биологические и социальные цели; в когнитивных процессах участвует не только мозг, но и тело, а понимание связано с действиями в среде, знания о которой содержатся в картине мира. В заключении статьи указываются тупики, трудности и опасности на пути к сильному искусственному интеллекту.

***Ключевые слова**: искусственный интеллект, машинное обучение, индукция, абдукция, понимание, когнитивная семантика, картина мира.*

***Цитирование:** Кузнецов О.П.* О машинном обучении, мифах о сильном искусственном интеллекте и о том, что такое понимание. *Онтология проектирования*. 2024. Т.14, №4(54). С.466-482. DOI: 10.18287/2223-9537-2024-14-4-466-482.

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Введение

Книга американского учёного и предпринимателя Э. Ларсона «Миф об искусственном интеллекте» [1] – хороший повод для обсуждения современного состояния искусственного интеллекта (ИИ) и его перспектив. Эта книга посвящена разоблачению мифов об ИИ, которые возникли ещё в середине прошлого века и активно обсуждаются в последнее десятилетие в научном (и не только) сообществе в связи с несомненными успехами нейросетевых технологий и машинного обучения (МО).

Прежде, чем начать рассказ об этой книге, следует договориться о словах.

Слова «искусственный интеллект» как в книге Ларсона, так и во всех обсуждениях проблем, связанных с ИИ, понимаются в двух смыслах: как компьютерная наука, разрабатыва-

ющая методы решения интеллектуальных задач, и как продукты этой науки, т.е. машинные реализации её методов. Иногда эти слова употребляются в обоих смыслах в пределах одного абзаца, и всякий раз понятно, что имеется в виду. Такова языковая реальность. Попытки многих профессионалов удержать только первое значение этих слов не устояли под напором огромного количества текстов, в которых «ИИ» понимается во втором смысле, и число которых постоянно увеличивается в связи с растущим интересом общества к проблемам ИИ.

Основной предмет обсуждения книги – ИИ во втором смысле, точнее, перспективы создания человекоподобного ИИ. Для этого понятия в английском и русском языках существует несколько синонимов. В книге Ларсона и в других англоязычных текстах человекоподобный ИИ фигурирует как *artificial general intelligence* (с соответствующей аббревиатурой *AGI*) или просто *general intelligence*. Из существующих русскоязычных синонимов (сильный, общий, универсальный, настоящий) можно использовать термин «сильный» как наиболее устоявшийся. Однако его естественный антоним «слабый» уместен лишь в качестве общей характеристики состояния интеллектуальных технологий. Неправильно называть конкретные интеллектуальные системы (ИС) «слабыми» только потому, что они решают (притом успешно) только одну задачу. Поэтому вместо «слабого ИИ» лучше использовать термин «ограниченный ИИ» (у Ларсона *narrow*), «узкий» или «узкоспециальный» ИИ.

1 О книге Э. Ларсона

1.1 История мифов об ИИ

Суть мифа об ИИ: возникновение человекоподобного ИИ, а в дальнейшем и сверхинтеллекта неизбежно, и оно произойдёт как бы само собой в результате постепенной эволюции систем ИИ.

Мифы об ИИ постоянно возникали, начиная с 1950-х гг. Впервые роковое слово «сингулярность» произнес Джон фон Нейман (1950-е гг., свидетельство Улама, приведённое в [2]): *«Постоянное ускорение технологического прогресса… свидетельствует о приближении некоей существенной сингулярности в истории человеческого рода, после наступления которой человеческая деятельность, какой мы её знаем, продолжаться уже не сможет».*

Джон Гуд, криптограф, работавший с Тьюрингом во время Второй мировой войны над расшифровкой немецких радиограмм, в 60-е годы утверждал, что, если машина способна развить интеллект человеческого уровня, то она сможет и превзойти обычное человеческое мышление, т.е. возникнет «ультраинтеллектуальная машина», которая «*сможет проектировать всё более совершенные машины, после чего непременно произойдет интеллектуальный взрыв, и человеческий интеллект останется далеко позади. Поэтому первая ультраинтеллектуальная машина станет последним изобретением, которое придётся создавать человеку*» [3].

В 1960-е годы многие ведущие специалисты по ИИ высказывались о скором создании ИИ, сравнимого с человеческим. Герберт Саймон в 1965 году предсказывал, что в ближайшие двадцать лет *«машины научатся выполнять любую работу, которая под силу человеку». В 1967 году Марвин Мински объявил, что «проблема создания ИИ будет по большей части решена в рамках одного поколения»*[1].

Начиная с 80-х гг. XX-го века, прогнозы развития ИИ стали приобретать апокалиптический характер: возникли предсказания о неизбежном создании сверхчеловеческого интеллекта, с появлением которого наступит «сингулярность». Типичным высказыванием в этом духе было заявление Вернора Винджа (Калифорнийский университет): «*В ближайшие трид-*

[1] Несколько позже (в 1975 г.) лидер советской кибернетики В.М. Глушков в интервью «Литературной газете» [4] заявил, что человекоподобный ИИ может быть создан до 2000 года.

цать лет у нас появятся технические возможности для создания сверхчеловеческого интеллекта. Вскоре после этого эра человечества подойдёт к концу… Я думаю, правильно будет назвать данное событие сингулярностью. Это та самая точка, где наши прежние модели перестают работать, и наступает новая реальность» [5].

Интересно, что примерно это же время характеризуется серией серьёзных неудач в сфере ИИ. Провалились проекты создания высококачественного машинного перевода, разработки диалоговых систем, способных пройти тест Тьюринга, проект «ЭВМ пятого поколения» и др. Интерес общества к проблемам ИИ и объёмы финансирования проектов, связанных с ИИ, сильно поубавились. И только появление Интернета, ставшего генератором больших массивов данных, и успехи МО, которые во многом основаны на работе с большими данными, дали новую жизнь этим пророчествам.

Одним из наиболее активных пропагандистов сингулярности стал неоднократно цитируемый в книге Ларсона Рэймонд Курцвейл — технический директор *Google* в области МО. В серии книг, опубликованных в 1990-2000-е годы [6-8], он развил концепцию, согласно которой технологии развиваются по экспоненциальному закону в соответствии со сформулированным им «законом ускорения отдачи». По его мнению, в 2029 году должен появиться полноценный человеческий ИИ, после чего к 2045 году возникнет сверхинтеллект [8]. Его рождение обозначит точку невозврата — сингулярность, с наступлением которой прогресс пойдёт по неизвестному пути. Это будет переломный момент, когда самыми разумными существами на планете станут машины.

Говорить о происходящей эволюции ИИ-систем в том же смысле, в каком говорят об эволюции в живой природе или об эволюции человечества, по крайней мере, некорректно. Эволюция некоторого сообщества агентов в строгом смысле слова предполагает способность этого сообщества к самовоспроизводству, т.е. к постоянной генерации агентами новых поколений, в ходе которой они наследуют свойства агентов прежних поколений, приобретают новые свойства и становятся всё более сложными. Никакого самовоспроизводства ИИ-систем не наблюдается. Более сложные ИИ-системы производятся не их предыдущими поколениями, а людьми. Самые совершенные на текущий момент системы являются всего лишь инструментами в руках человека. Поэтому так называемая «эволюция ИС» – это в действительности эволюция человеческих идей.

Впрочем, у фон Неймана (в отличие от Гуда, Курцвейла и их последователей) были основания размышлять об эволюции машин, поскольку он всерьёз рассматривал возможности создания самовоспроизводящихся автоматов и их эволюции в указанном выше строгом смысле [9]. Его идеи имели продолжение (см., например, [10]), но в процессе развития ИС реализация этих идей практически отсутствует. Решится ли когда-нибудь человечество заняться созданием самовоспроизводящихся машин в промышленном масштабе и тем самым выпустить очередного (вслед за атомной бомбой и генетическим редактированием) джина из бутылки - вопрос будущего.

Типичным представителем более умеренной позиции при обсуждении проблем создания сильного ИИ является известный специалист в области ИИ Стюарт Рассел, который «*стремится различать серьёзную работу в области разработки сильного ИИ и его изображение в массовой культуре*». Он и его сторонники не сомневаются в скором появлении сильного ИИ. Предвидя серьёзные проблемы в отношениях сильного ИИ и человека, Рассел предлагает «*действовать на опережение и встраивать в сверхинтеллектуальные машины будущего определённые принципы*», которые должны исключить возможность апокалиптических сценариев. Эти принципы подробно изложены в его книге [11].

Заметим, что эти принципы своей декларативностью и внутренними противоречиями напоминают знаменитые «три закона робототехники» Айзека Азимова.

Популярность мифов об ИИ в обществе Ларсон объясняет тем, что они имеют признаки китча – продукта массовой культуры. «*Во-первых, китч подразумевает упрощение сложных идей. Необходима простая и понятная история. Во-вторых, китч предлагает лёгкие решения, избавляющие людей от волнений по поводу жизненных проблем, вместо того чтобы решать их при помощи серьёзного, тщательного обсуждения. Отличный тому пример — мечтательное представление о том, что когда-нибудь появится невероятный андроид, обладающий сверхинтеллектом, который*

перестроит человеческое общество с его устаревшими традициями и взглядами, и люди вступят в новую эру, где, по счастью, не будет места былым спорам о Боге, природе сознания, свободе воли, праведной жизни и тому подобном» [1].

Критика мифов об ИИ ведётся в книге по двум направлениям: научному и социальному. По мнению Ларсона, с научной точки зрения прогноз о неизбежности появления сверхразума не имеет под собой оснований: современные тренды и методы построения ИС не ведут не только к возникновению сверхчеловеческого интеллекта, но даже и к созданию сильного ИИ, т.е. ИИ, сравнимого с человеческим. С социальной точки зрения эти мифы небезобидны и ведут к серьёзным негативным последствиям для развития науки и для общества в целом.

1.2 ИИ: логический вывод, представление знаний и машинное обучение

Обоснование недостижимости сильного ИИ методами МО Ларсон строит на подробном обсуждении проблем, связанных с логическим выводом и его использованием в ИС.

Важнейшим аспектом интеллекта – как естественного, так и искусственного – является умение рассуждать, т.е. делать умозаключения. «*Если ИС совершенно не способна строить умозаключения, то она не заслуживает того, чтобы называться интеллектуальной*» [1]. В логике известны три основных вида умозаключений (логических выводов): дедукция, индукция и абдукция.

Дедукция – вывод от общего к частному: *если A, то B; A истинно; следовательно, B истинно.*

Индукция – вывод от частного к общему: *все известные объекты класса A обладают свойством B; следовательно, все объекты класса A обладают свойством B.* Важным частным случаем индукции является статистический вывод. Пример: *в некоторой выборке элементов класса A x процентов обладают свойством B; следовательно, во всем классе A x процентов его элементов обладают свойством B.*

Абдукция - вывод от следствия к причине: *если A, то B; B истинно; следовательно, A истинно*, т.е. *A* – причина *B*.

Каждый из этих видов имеет свои достоинства и недостатки. Дедуктивный вывод является достоверным: если обе его посылки истинны, то заключение гарантированно истинно. Поэтому он служит хорошим средством для обоснования высказанных утверждений и проверки возможных ошибок в рассуждениях. Однако он не создаёт новых знаний в том смысле, что не порождает новых общих утверждений. Кроме того, формальная дедукция не гарантирует релевантности, так как истинная посылка вида «если…то» не всегда описывает причинно-следственную связь. Индукция даёт новые знания, т.е. порождает общие утверждения, но не гарантирует их достоверность. Никакое число положительных примеров не гарантирует, что следующий пример не окажется отрицательным. «*Наши наблюдения и тесты всегда неполны. Корреляции могут указывать на достоверную реальную причину (как на некоторое знание), но мы могли что-то упустить при наблюдении и проверке того, что случилось. Корреляция может быть мнимой или случайной. Возможно, мы ищем не то. Выборка может быть слишком мала или нерепрезентативна по причинам, которые выяснятся лишь позднее*» [1].

> Индуктивный вывод о будущем характере процесса, развивающегося во времени, – это экстраполяция, предполагающая, что закономерность развития, наблюдаемая в прошлом и настоящем, сохранится и в будущем. Без дополнительных аргументов такой вывод не является обоснованным: например, известно немало процессов с длительным степенным или экспоненциальным ростом, которые со временем выходят на плато. Именно такой необоснованной экстраполяцией является «закон ускорения отдачи» Курцвейла.

Первые успехи в автоматизации дедуктивного вывода были достигнуты на заре ИИ. Уже в конце 50-х гг. XX-го века была создана программа «Логик-теоретик» А. Ньюэлла, К. Шоу и Г. Саймона, которая могла доказывать многие логические теоремы. Но для того, чтобы делать содержательные выводы, нужно иметь обширную базу знаний (иначе откуда брать истинные

посылки для вывода?). Поэтому в 60-70-е годы в ИИ одно из первых мест по числу исследований заняла инженерия знаний - методы извлечения, представления, хранения знаний, доступа к ним и работы с ними. Довольно быстро стало ясно, что главная проблема заключается не столько в разработке методов, сколько в необъятности тех обыденных знаний, которые хранятся в голове у человека и которые следовало бы хранить в любой ИС, претендующей на человекоподобное поведение.

«*Министерство обороны США в своё время вкладывало огромные суммы в создание крупных баз обыденных знаний. Эксперты, компетентные в логике и вычислениях, по капле скармливали этим системам тривиальные утверждения, например, «*у живого человека есть голова*», или «*поливалки брызгаются водой*», или «*от воды вещи мокнут*» и так далее*» [1].

Это занятие превратилось в «*проблему бездонного ведра: задача наполнения вычислительной базы знаний утверждениями, выраженными в виде логических высказываний, оказывается бесконечной. Не удаётся решить даже простейшие задачи, основанные на здравом смысле, — например, рассуждать о происшествиях, возникающих в городском квартале или районе, — если не будут реально кодифицироваться огромные объёмы, казалось бы, нерелевантных знаний*» [1].

Уже более 30 лет существует проект *Cyc* Дугласа Лената. Система *Cyc* содержит множество логических утверждений о фактах и общих понятиях типа «Объект не может находиться более чем в одном месте в одно время», а также алгоритмы вывода из этих утверждений. «На лекции в 2015 году Ленат сказал, что в настоящий момент в *Cyc* содержится 15 миллионов утверждений и предположил, что, вероятно, это около пяти процентов от необходимого» [12]. Тем не менее, этот проект не оказал существенного влияния на основные направления исследований в сфере ИИ.

МО индуктивно; проблемы, характерные для индуктивного вывода, относятся и к МО. Главная из них – недостоверность индукции. «*Границы мира системы МО строго ограничены тем набором данных, на котором она тренируется. В реальном мире наборы данных генерируются круглосуточно, семь дней в неделю. Следовательно, любой конкретный набор данных охватывает лишь очень небольшой период и лишь частично отражает свойства, присущие системам в реальном мире*» [1].

Взрослый человек в некотором смысле «обучен всему». В огромном множестве разнообразных ситуаций, не выходящих за рамки «обычного», т.е. согласующихся с накопленным опытом, разные люди ведут себя по-разному, но, как правило, адекватно – отклонения от нормы довольно быстро отмечаются окружением. В любой необычной ситуации какие-то частицы накопленного опыта (начиная с базовых знаний типа упомянутых выше: «объект не может находиться более чем в одном месте в одно время» и т.д.) всё равно оказываются полезными; с их помощью (и – что крайне важно! – с помощью неразгаданного пока наукой «здравого смысла») человек постепенно начинает в этой ситуации как-то ориентироваться.

Обученные системы не таковы. «*Если будущее представляется неопределённым, а изменения желательны, системы приходится переобучать. Машинное обучение может двигаться лишь за потоком нашего опыта, имитируя полезные (остаётся на это надеяться) регулярности. При этом нас ведёт именно разум, а не машина*» [1]. Другими словами, не система решает, что ей надо переобучиться, а человек, который её обучает.

А что же с абдукцией? В книге [1] абдуктивному выводу и его первооткрывателю Чарльзу Сандерсу Пирсу уделяется значительное место. Абдукция – это рассуждение от события к его причинам. Недостоверность абдукции ещё более очевидна, чем недостоверность индукции. То, что из A следует B и событие B произошло, вовсе не означает, что причиной B является именно A: возможных причин события B может быть много. Иначе говоря, могут быть истинными посылки «если A_1, то B», …, «если A_n, то B», причём n неопределённо велико, A_1, …, A_n – возможные гипотезы о причине, и только некоторые из них, быть может, верны. При этом в случае, когда верны A_i, B и «если A_i, то B», то A_i может оказаться не причиной B, а корреляцией, т.е. A_i и B могут быть следствиями одной и той же причины A_j.

Именно с множественностью возможных причин связаны проблемы врача, ищущего причины боли в желудке, проблемы инженера, выясняющего, почему прибор не работает, и проблемы детектива, занятого поиском преступника и мотивов преступления. Для успеха этого поиска нужно, во-первых, иметь знания о возможных гипотезах, причём, если врач эти знания либо черпает из своего опыта, либо получает из медицинской литературы, то для детектива в каждом преступлении много уникального, и соответствующие знания (улики, сведения о подозреваемых и т.д.) надо ещё добывать. Во-вторых, гипотезы надо проверять; при этом многие гипотезы человеком отбрасываются практически сразу, как не имеющие отношения к делу. Умение отличать существенное от несущественного, релевантное от нерелевантного – важная черта человеческого интеллекта.

Выводы, которые мы совершаем, зачастую являются догадками, которые кажутся нам релевантными или правдоподобными, — а не дедукцией или индукцией. Если дедукция и индукция не подходят, то у нас обязательно должна быть теория абдукции. Поскольку у нас её (ещё) нет, можно сделать вывод, что пока мы не вышли на путь к сильному ИИ [1].

На абдукцию Ларсон возлагает особые надежды. Однако вряд ли они оправданы. Как видно из аргументов, изложенных в книге, решение проблемы абдукции неразрывно связано с формализацией названных выше свойств интеллекта, которые обобщённо называют здравым смыслом. А до этой формализации ещё далеко, и, как считает Ларсон, МО, весь датацентричный ИИ, к ней не приближает. ***Путь к сильному ИИ через машинное обучение – это тупик.***

Ещё одно направление в современном ИИ, рассматриваемое сторонниками мифов как путь к сильному ИИ – это обработка естественного языка. Этот путь в книге [1] также подвергается серьёзному анализу и критике. Главным аргументом здесь является отсутствие понимания.

За последнее десятилетие машинное обучение и большие данные позволили добиться существенного прогресса при решении некоторых задач, но, как правило, при помощи путей, которые позволяют обойтись без фактических знаний и понимания [1].

Основное внимание в этой части книги уделяется вопросно-ответным системам – в частности, потому, что с ними связаны многочисленные попытки пройти тест Тьюринга. Обсуждаются предложения по его усовершенствованию: в частности, ужесточение требований к участникам теста с целью исключить различные трюки (приёмы, известные со времен «Элизы»: повтор вопроса в качестве ответа, общие фразы, не отвечающие на вопрос, попытки уклонения от ответа и т.д.), а также методы разработки вопросов, имеющих целью поймать систему на непонимании смысла. Особый интерес представляют предложенные Г. Левеском и его коллегами [13, 14] «схемы Винограда» – вопросы, содержащие неоднозначность, разрешить которую можно, только используя либо контекст, либо некоторые знания о предметах, упомянутых в вопросе.

На момент выхода книги [1] лучшие результаты вопросно-ответных систем по ответам на схемы Винограда не превышали 62% [1]. Вот как объясняет Ларсон причины этих неудач.
«…любые два слова или фразы, заключённые в одном вопросе, резко снижают ожидаемую частоту встречаемости такого сочетания. Следовательно, в больших данных все эти примеры относительно редки, хотя и просты. А в случаях, когда два имени или именных словосочетания все-таки могут встречаться в сети, достаточно просто изменить отношение между этими существительными, переставив их местами. Этот способ позволяет сбить частотность и победить современные методы, ориентированные на большие данные. Поэтому вопросы из схем Винограда во всех отношениях не поддаются машинной имитации — именно этим и объясняются слабые результаты тех систем, что применялись для автоматизации прохождения этого теста» [1].

Ещё один плохо формализуемый аспект здравого смысла — прагматика, т.е. понимание целей и намерений говорящего, которые передаются через контекст, жесты и интонации.

Например, вопрос за столом «Можешь передать мне соль?» — это вовсе не вопрос, а просьба, которая ожидает не ответ «Могу», а действие, т.е. передачу соли.

Подводя итог обсуждению успехов и неудач машинного обучения, Ларсон подробно рассказывает о том, как создавалась система *Watson*, которая на момент выхода книги была наиболее успешной вопросно-ответной системой. Отдавая должное изобретательности команды её разработчиков, он заключает, что «*Watson также не был шагом в развитии ИИ, а только лишний раз подтвердил, что поиск сильного ИИ по-прежнему вязнет в путанице и тайнах. Хотя команда IBM действительно добилась впечатляющей победы, воспользовавшись мощной гибридной системой, но эта работа не помогла подобрать ключ к пониманию языка»* [1].

Справедливости ради заметим, что в 2021 году, когда вышла книга Ларсона, ещё не были в центре общественного внимания большие языковые модели (*LLM*; наиболее известны модель *BERT* и серия моделей *GPT*). В книге [1] они не обсуждаются. Их возможности ещё предстоит выяснять, хотя уже ясно, что они гораздо выше возможностей предыдущих систем обработки языка. Однако они тоже основаны на обучении и больших данных, т.е. имеют индуктивную (точнее, статистическую) природу, а потому вся критика, связанная с недостатками индукции и отсутствием понимания, относится и к ним.

1.3 Социальные аспекты мифов об ИИ: наука и общество

Мифы об ИИ небезобидны и имеют нежелательные социальные последствия. Вера в миф о саморазвивающейся эволюции ИС, по мнению Ларсона, ослабляет веру в человеческий потенциал. «*В нынешней мифологии человеческий разум начинает восприниматься как устаревающая модель грядущих машин*» [1]. В основе этой мифологии лежат растущие вычислительные мощности, генерируемые Интернетом Большие данные и ИИ, понимаемый как машинное обучение на этих данных.

На заре Интернета, предоставившего невиданные ранее возможности для общения и обмена информацией миллионам людей, преобладали пророчества о всплеске человеческого потенциала. «*Веб не только сулил сделать нас умнее и осведомлённее, но и должен был помочь нам эффективнее сотрудничать, чтобы мы выстраивали современные цифровые пирамиды, преобразовывали науку и культуру*». Эти иллюзии довольно быстро сменились «*мировоззрением, в котором люди рассматриваются как винтики в гигантской машине»* [1].

В 2008 г. редактор журнала *Wired* Крис Андерсон заявил, что в связи с появлением Больших данных приходит конец теоретической науке [15]. В 2015 году Шон Хилл (один из руководителей проекта *Human Brain Project - HBP*) указывал, что будущее науки связано с крупными коллаборативными проектами, а отдельных учёных лучше всего расценивать как единицы «роя» [16]. Инициатор этого проекта Генри Маркрам утверждал, что такие гении, как А. Эйнштейн, теперь уже не нужны. «*Нас тормозит всеобщее мнение, что нужен новый Эйнштейн, который объяснил бы, как работает мозг. На самом же деле нам требуется отодвинуть в сторону собственное эго и создать новый вид коллективной нейронауки*» [17]. Ларсон отмечает, что эти взгляды приводят «*к компьютеро-центричному мировосприятию, где человеческий потенциал принижается в пользу господства машин. Наука движется вслед за онлайновой культурой, от человеческих идей до мегатехнологий, к консолидации власти в крупных технологических компаниях и к общей стагнации и замедлению инноваций*» [1].

Проект *HBP* первоначально предполагал построение цифровой копии мозга на основе данных обо всех нейронах и синапсах мозга. Его целью по мысли Маркрама было движение «*от генетического, молекулярного уровня к нейронам и синапсам, далее к цепям нейронов, макроцепям, мезоцепям, долям мозга — до тех пор, пока* ***не возникнет понимание*** *того, как связаны между собой все эти уровни и как они определяют поведение и формируют сознание*». Предполагалось, что по мере накопления данных о мозге понимание его работы возникнет как бы само собой. При этом исследования «верхнего уровня» (когнитивные архитектуры, мышление и поведение) оттеснялись на второй план. Эта политика вызвала серьёзную критику, в результате ко-

торой содержание проекта свелось к разработке инструментов и методов моделирования мозга. Проект просуществовал 10 лет и был завершён в 2023 г., не достигнув объявленной цели.

Вот как описала достижения *HBP* экспертная комиссия, которая оценивала итоги проекта, в своём пресс-релизе: «Уже сегодня инфраструктура *EBRAINS* открывает возможности для новых приложений в области здоровья мозга и технологий, производных от мозга. *HBP* установила новую парадигму цифровой нейробиологии и новую междисциплинарную культуру сотрудничества. Среди особо важных достижений - ведущие цифровые атласы мозга, передовые платформы для моделирования мозга во всех масштабах, применение когнитивного моделирования и персонализированной медицины, а также выдающиеся достижения в области нейроморфных вычислений, нейро-вдохновлённой робототехники и ИИ».

Неплохо, но всё же это очень далеко от обещанной «точной модели мозга».

В качестве примера успешного проекта, основанного на современных технологиях, Ларсон приводит открытие бозона Хиггса. «*Случай Хиггса особенно впечатляет в плоскости теории, а не только в плоскости больших данных… Питер Хиггс предвосхитил открытие этой частицы в 1964 году; БАК (Большой Адронный Коллайдер) впоследствии лишь подтвердил её существование. Это пример правильного использования технологий, дополняющих человеческие озарения*» [1].

В итоге: «*порочная привычка продвигать ИИ больших данных в качестве панацеи угрожает прогрессу в фундаментальных дисциплинах, таких как нейронауки, — несмотря на смелые заявления Маркрама и других энтузиастов. В данном случае вывод таков: миф самым реальным образом сказывается на человеческом будущем и на реальной науке»* [1].

Ещё один аспект социальных последствий мифов об ИИ - индустриализация науки. «*Технологические стартапы, ещё недавно будораживишие инвесторов, теперь сводятся к идее «пусть их купит технический гигант вроде Google…» Эти гиганты ... монополизировали доступ к инновациям, так как ИИ больших данных всегда лучше работает у тех, у кого в наличии больше данных*» [1]. Ещё Норберта Винера беспокоила эта тенденция, которую он в своей опубликованной только после его смерти статье [18] назвал «мегабаксовой наукой». Опасность заключается, во-первых, в том, что новые идеи в своей начальной стадии не гарантируют прибыли, и потому инвестировать в них рискованно. Во-вторых (и здесь он предвидел современные мысли о «роевом разуме»), «*расчёт на генерацию по-настоящему значительных новых идей путём умножения низкокачественной человеческой деятельности и случайной перестановки имеющихся идей без участия первоклассного разума, который руководил бы отбором этих идей — очередная версия истории о печатающих обезьянах*» [18].

Кратким итогом обсуждения в книге [1] перспектив развития ИИ при сохранении современных трендов можно считать следующую цитату: «*Мы фактически наблюдаем эволюцию подвида индуктивного ИИ, хорошо функционирующего в узких основанных на больших данных окружениях, но однозначно неспособного усваивать здравый смысл и достигать подлинного понимания. Этот подход не имеет ничего общего с сильным ИИ*» [1].

В заключение раздела отметим ещё одно препятствие на пути к сильному ИИ, которое до сих пор сравнительно мало обсуждается. В книге Ларсона о нём не говорится. Речь идёт о непомерном росте вычислительных мощностей (и соответственно, оборудования), энергопотребления и экологических последствий (выбросы CO_2, расход воды для охлаждения и т.д.) при эксплуатации больших ИС. Например, в обзоре [19] отмечается, что «*вычисления, требуемые для обучения, должны расти, по крайней мере, как полином четвёртого порядка от качества... Из-за сложности глубокого обучения качество может быть значительно хуже*». А ведь речь идёт об ограниченных системах, решающих какую-то одну задачу. И в этом отношении ИИ проигрывает мозгу, который весит около 1 кг и потребляет мощность примерно 20-30 Вт.

Несмотря на успехи ИИ, базовые преимущества мозга перед компьютером сохраняются. Следующие прорывы в ИИ невозможны без прорывов в изучении информационных процессов мозга. И среди главных проблем на этом пути – проблема понимания.

2 О понимании

Люблю обычные слова,
Как неизведанные страны.
Они понятны лишь сперва,
Потом значенья их туманны.
Их протирают, как стекло,
И в этом наше ремесло.
Давид Самойлов «Слова» (1961)

«Понимание» – одно из ключевых слов при обсуждении перспектив и проблем развития ИИ, а «отсутствие понимания» – один из главных аргументов в критике мифов о скором появлении человекоподобного ИИ. Об этом много говорится в книге Ларсона. Мелани Митчелл посвятила несколько глав своей книги [12] «барьеру понимания» как важному препятствию на пути к сильному ИИ. А Генри Маркрам в проекте *HBP* рассчитывал на то, что создание цифровой копии мозга автоматически приведёт к пониманию того, как он работает. О том, что её построение не привело бы к пониманию работы мозга, говорит эксперимент, описанный в статье [20], который имеет непосредственное отношение к проблеме понимания.

Нейробиологи, вооружившись методами, обычно применяемыми для изучения живых нейроструктур, попытались использовать их, чтобы понять, как функционирует простейшая микропроцессорная система. Объектом исследования стал чип (*MOS* 6502), использованный во множестве персональных компьютеров и игровых приставок. Об этом чипе известно всё, но исследователи сделали вид, что не знают ничего, и попытались понять его работу, изучая теми же методами, которыми изучают живой мозг.

Была удалена крышка, под оптическим микроскопом изучена схема с точностью до отдельного транзистора. Чип был подвергнут тысячам измерений одновременно: во время его работы измерены напряжения на каждом проводке и определено состояние каждого транзистора. Это породило поток данных в полтора гигабайта в секунду, который анализировался. Строились графики всплесков от отдельных транзисторов, выявлялись ритмы, отыскивались элементы схемы, отключение которых делало её неработоспособной, находились взаимные зависимости элементов и блоков и т.п. Однако *понять*, что делает этот микропроцессор, т.е. какова его функция, исследователям так и не удалось.

В статье [20] сформулирован следующий критерий понимания: *«Понимание части системы возникает, когда можно описать её входы, выходы и преобразования от входов к выходам настолько точно, что эту часть можно заменить искусственным компонентом».*

Этот критерий можно истолковать в двух вариантах: сильном и слабом. Сильный вариант означает, что для понимания нужна математическая модель этой части системы – именно математическая, а не квадратики со стрелками, – поскольку только математика умеет точно описывать преобразования. Слабый вариант предполагает, что для ограниченного числа входных воздействий известно, какие выходные реакции последуют. С этим вариантом имеет дело обычный пользователь бытового прибора, который внимательно прочёл инструкцию и знает, какие кнопки надо нажимать, чтобы получить нужные действия. Он понял, *как* прибор работает, и даже – как заменять некоторые компоненты (например, батарейки); но не понимает, *почему*: полное описание преобразований (т.е. как прибор устроен внутри и каковы причины, по которым происходят те или иные его действия), он не знает – да ему это и не нужно.

Описанный в [20] эксперимент иллюстрирует важный методологический принцип: *интерпретация данных не может проводиться в терминах самих данных.* Для неё нужен язык более высокого уровня с другой системой понятий. Для науки это язык теории. Как сказал А. Эйнштейн, *«Лишь теория решает, что мы ухитряемся наблюдать».* Ключевое слово – *интерпретация*, иначе говоря, осмысление, придание значения. В науке интерпретация фактов, т.е.

их понимание происходит в терминах теории. В обыденном мышлении человека понимание происходит в терминах его *картины мира*.

Понятие картины мира возникло давно. В обширном обзоре истории этого понятия, содержащемся в книге [21], отмечается, что впервые оно появилось ещё в XVIII веке, а систематически стало употребляться во второй половине XIX века для описания состояния объективного, т.е. общенаучного знания о мире. Идея субъективной картины мира («*Umwelt*»), присущей каждому живому существу, впервые была высказана немецким биологом Якобом фон Икскюлем в 1909 г. [22][2]. Интерес представляют субъективные, когнитивные аспекты картины мира человека, которые следовало бы назвать когнитивной семантикой. В настоящее время под когнитивной семантикой, как правило, подразумевается раздел лингвистики, который исследует языковые средства передачи различных смыслов. Следуя подходу Дж. Лакоффа [25], будем считать когнитивной семантикой исследование всех (не только языковых) средств извлечения, хранения смыслов и выражения их в языке, жестах, поведении и т.д.

В наше время «хайпов», когда на каждую перспективную идею набрасываются сотни исследователей, а публикации более чем пятилетней давности считаются устаревшими, при этом существуют концепции, которые ждут своего часа десятилетиями. Такова замечательная книга Лакоффа «Женщины, огонь и опасные вещи» [25], которую цитировали такие известные специалисты, как нейропсихолог Антонио Дамасио и нейролингвист Стивен Пинкер. Однако её идеи по-прежнему остаются (незаслуженно!) в стороне от мирового тренда.

Концепция когнитивной семантики, которая излагается ниже, основана на работах Дж. Лакоффа [25], С. Пинкера [26][3], А. Дамасио [27], А. Сета [28] и заключается в следующем.

1. Понимание – это интерпретация в терминах картины мира человека[4].
2. Картину мира строит наш мозг, и эта картина (например, интуитивная физика: то, как человек представляет себе внешний мир) может сильно отличаться от того, каким мир является на самом деле, т.е. каким его представляет научное знание.
3. Картина мира структурируется через категоризацию опыта человека, в ходе которой формируются понятия и связывающие их образно-схематические структуры.
4. Значения (смыслы) формируются раньше, чем формируются понятийные структуры: они возникают из нашего допонятийного телесного опыта. Смыслы первичны, их языковое оформление вторично.
5. В основе значений лежат биологические и социальные цели, в первую очередь, выживание.
6. В когнитивных процессах участвует не только мозг, но и тело.
7. Понимание связано с действиями в среде, знания о которой содержатся в картине мира.

Краткие комментарии к этим тезисам.

1. Следует отличать отсутствие понимания от неправильного понимания. Отсутствие понимания означает, что воспринимаемое не удаётся вписать в имеющуюся картину мира: у человека нет нужных понятий (и, соответственно, слов), чтобы его описать или задать разумные вопросы. В этом случае картину мира приходится достраивать, т.е. приобретать новые знания.

Неправильное понимание означает, что нужные понятия нашлись и интерпретация произошла, но она не соответствует реальному миру. Здесь возможны два варианта. Первый: противоречие с картиной мира слишком явно; возникает когнитивный диссонанс и желание его устранить [31]: либо интерпретация отвергается («Этого не может быть!», «Я не могу этого представить!» и даже «Я этого не понимаю!» - хотя на самом деле «Понимаю, но не принимаю»), либо картину мира приходится перестраивать (например, изменять своё мнение

[2] Подробнее о взглядах фон Икскюля см. [23, 24].
[3] Обзоры книг Лакоффа и Пинкера содержатся в статьях [29] и [30], соответственно.
[4] Указанные авторы, если и употребляют понятие «картина мира», то как метафору, а не как термин. Здесь это понятие является ключевым.

о каком-то знакомом человеке). Второй: явное противоречие сразу не обнаруживается, и человек долгое время (может быть, всю жизнь) с ним живёт; например, считает, что Солнце вращается вокруг Земли или что все лебеди белые. Известно, что обыденное мышление склонно сохранять существующую картину мира и сопротивляться её перестройке. В [31] отмечается, что при наличии двух альтернатив, одна из которых соответствует картине мира, а другая противоречит ей, человек предпочитает искать дополнительные аргументы в пользу первой альтернативы. *«Человек будет искать такие источники информации, которые способствовали бы добавлению консонантных элементов, и будет избегать источников, увеличивающих диссонанс»* [31]. Поскольку перестройка картины мира требует значительных когнитивных (а иногда и эмоциональных) усилий, стремление к её сохранению вполне соответствует концепции «когнитивной лени» Д. Канемана [32], т.е. минимизации когнитивных усилий.

О стремлении к устойчивости картины мира, различиях между критическим, обыденным и догматическим типами мышления, а также о том, чем явное противоречие отличается от неявного, подробно говорится в [33].

2. Базовая картина физического мира строится мозгом на основе перцептивного опыта, который, как указывает А. Сет [28], *«определяется содержанием нисходящих предсказаний, а не восходящих сенсорных сигналов»*. Иначе говоря, мозг постоянно предсказывает свои ощущения и проверяет свои предсказания опытным путём, обучаясь на ошибках. *«Мы воспринимаем окружающий мир, чтобы эффективно в нём действовать, добиваться своих целей и в конечном итоге повышать свои шансы на выживание. Мы воспринимаем мир не таким, какой он есть, а таким, каким он нам полезен»* [28]. Поэтому «интуитивная физика», т.е. система представлений обычного человека о физическом мире, сильно отличается от научных физических знаний. Уже *«классическая ньютоновская физика глубоко контринтуитивна»* [26], не говоря о квантовой механике и теории относительности. Тем не менее в стандартных ситуациях этой интуитивной физики оказывается достаточно.

Типичный пример – слесарь-электрик, который не знает ни уравнений Максвелла, ни теории цепей Кирхгофа и представляет электричество как поток некой жидкости. Тем не менее он прекрасно справляется с ремонтом бытовых электроприборов и домовых электрических сетей. Это случай «слабого» понимания, о котором говорилось при обсуждении статьи [20]. Он хорошо иллюстрирует тезис о том, что обыденному мышлению нужна не истина, а польза.

Истина и польза не являются альтернативами. Но истина не входит в число неотъемлемых жизненных целей человека. На неё, как правило, нужно тратить когнитивные усилия, размышлять, а действовать, чаще всего, нужно здесь и сейчас. Поэтому в стандартных ситуациях, которые составляют бо̀льшую часть жизни человека, действует быстрая Система 1 Канемана [32], основанная на схемах. Но и медленная, рациональная Система 2 не всегда добирается до истины: либо по объективным причинам (если получение истины слишком сложно), либо из-за недостаточных знаний, либо по причине когнитивной лени [32, 34]. Если приходится выбирать между истиной и пользой, человек, как правило, выбирает (иногда вынужденно) пользу – можно вспомнить историю Галилея, которому пришлось выбирать между истиной и выживанием.

Сказанное относится только к обыденному мышлению. Целью научного мышления является получение истины. В промежуточном положении находятся различные формы профессионального мышления: не случайно доказательная медицина – это лишь часть медицины.

3. Структура картины мира описана в книге Лакоффа [25]. Она имеет два уровня: а) базовый, допонятийный, определяемый гештальтным восприятием и сенсорно-двигательным опытом; б) абстрактные понятийные структуры. Категории базового уровня – гештальты (целостно воспринимаемые образы); отношения между ними строятся на основе образных схем[5]

[5] Важную роль схем в организации знаний и рассуждений отмечают многие исследователи когнитивных процессов человека. Наряду с упомянутыми книгами [25, 26, 32] следует назвать книги [35, 36]. Важность схем отмечал М. Минский, назвав их фреймами [37]. Помимо указанных базовых схем человек обладает огромным количеством схем, сформированных в личном опыте (профессиональные схемы, схемы типичных ситуаций и т.д.).

типа *вместилище, путь, связь, сила, равновесие, верх-низ, спереди-сзади, часть-целое, центр-периферия*, которые часто встречаются в нашем телесном опыте. Именно со структур базового уровня, воспринимаемых непосредственно, начинается формирование картины мира у детей. Абстрактные понятийные структуры возникают либо в результате операций обобщения-конкретизации, либо с помощью метафорического переноса структур базового уровня на абстрактный уровень. Например, категория времени характеризуется схемой спереди-сзади (будущее впереди, прошлое позади). При этом категории базового уровня находятся в середине иерархии общего-конкретного. Обобщение происходит вверх от базового уровня, конкретизация – вниз. Пример: собака – базовая категория, хищник – обобщение, овчарка – конкретизация. Эту структуру категорий впервые описала Элеонора Рош в своей теории прототипов [38].

В рабочей (оперативной) памяти человека всегда присутствует только незначительная часть картины мира, и только к этой части предъявляются требования согласованности и непротиворечивости. В этом – одна из главных причин типичной для обыденного мышления фрагментарности и слабой чувствительности к противоречиям. Часто фрагмент картины мира, находящийся в рабочей памяти, называют репрезентацией (например, в [36]).

Репрезентации – это конструкции, зависящие от обстоятельств. Они построены в конкретном индивидуальном контексте для специфических целей: для осведомлённости в данной ситуации, для того, чтобы быть готовым к требованиям текущей ситуации и понимать текст, инструкцию, проблему. Репрезентации учитывают всю совокупность элементов ситуации или задачи. Они очень специфичны, детализированы и непрочны. Репрезентация модифицируется, если изменилась вся ситуация или незаметный элемент вдруг стал заметным.

Знания – это конструкции, обладающие постоянством и существенно не зависящие от выполняемой задачи. Знания хранятся в долговременной памяти [36].

4. Значения (смыслы) для человека первичны и возникают раньше понятий и языка. Например, годовалые дети, ещё не владеющие языком, уже имеют некоторый набор базовых значений, который они получают из своего телесного опыта. Картина мира содержит только значимые понятия, и люди оперируют только теми понятиями, которые включены в их картину мира. Сами *понятия хранятся в форме, гораздо более абстрактной, чем предложения. ... люди плохо помнят конкретные предложения, из которых они почерпнули свои знания. Однако это не мешает людям запоминать суть того, что они услышали или прочитали* [26]. При этом «суть» воспринятого у разных людей может отличаться, поскольку у них разные картины мира и, соответственно, разные интерпретации.

Часто можно наблюдать, как человек ищет слова, чтобы выразить свою мысль. Вот что говорит А. Эйнштейн о том, как он думает: *«Слова, написанные или произнесённые, не играют, видимо, ни малейшей роли в механизме моего мышления. Психическими элементами мышления являются более или менее ясные знаки или образы, которые могут быть «по желанию» воспроизведены или скомбинированы. ... Элементы, о которых я только что говорил, у меня бывают обычно визуального или изредка двигательного типа. Слова или другие условные знаки приходится подыскивать (с трудом) только во вторичной стадии, когда эта игра ассоциаций дала некоторый результат и может быть при желании воспроизведена»* [39].

5. Значимость воспринимаемых фактов и событий определяется их возможным влиянием на достижение тех или иных целей. Человек – это прежде всего биологический организм, и одна из его фундаментальных целей – выживание, т.е. пребывание в одном из состояний, совместимых с жизнью. Поскольку человек ещё и социальный организм, то под выживанием имеется в виду не только поддержание физиологических параметров, совместимых с жизнью, но и таких социальных параметров, как качество жизни, «выживание в обществе», т.е. занятие в нём определённого места, обеспечивающего это качество, и т.д. *«Последствия достижения или недостижения сложной социальной цели способствуют (или воспринимаются как способствующие), хотя и косвенно, выживанию и качеству выживания»* [27].

6. В когнитивных процессах участвует не только мозг, но и тело. *Эмоции, чувства и биологическая регуляция – все они играют роль в рассуждениях человека. Низшие слои нашего организма включены в цикл высших рассуждений* [27]. Об этом говорят все авторы книг [25-28]. Дж. Лакофф называет наше мышление воплощённым (*embedded*), т.е. непосредственно связанным с телом. Книга известного нейропсихолога А. Дамасио [27] целиком посвящена обоснованию этого тезиса. Характерно её название: «Ошибка Декарта» (известно, что Декарт резко отделял разум от тела). Телесные состояния и механизмы напрямую влияют на когнитивные процессы. *«Мозг, отделённый от тела, не может иметь нормальный разум»* [27].

7. Картина мира нужна для того, чтобы действовать в мире, опираясь на знания о нём. На это указывал ещё Икскюль [22]. Действия нужны не только для достижения своих целей, но и для проверки (и, возможно, коррекции) картины мира. Мозг постоянно предсказывает свои ощущения и проверяет эти предсказания опытным путём. *«Основополагающее занятие мозга – порождение действий и постоянная их калибровка с учётом сенсорных сигналов. С этой точки зрения мозг предстает динамичной, активной системой, непрерывно прощупывающей среду и изучающей последствия»* [28]. При этом действие не обязательно должно быть физическим. В социальной среде действие – это и коммуникация: задавание вопросов, способствующих пониманию, навязывание своей картины мира собеседнику или обществу и т.д.

Из изложенных тезисов вырисовывается примерный образ понимающей системы: это автономный активный агент, действующий в среде, умеющий ставить цели, стремиться к их достижению и способный формировать и корректировать свою картину мира. Читатель скажет: ведь это робот! Да, современные роботы автономны, активны, умеют достигать своих целей и даже создавать свою картину мира (например, пространства, в котором они передвигаются). Более того, у них уже есть нечто, что можно интерпретировать как эмоции и темперамент [40]. Но цели им ставит человек, и только под эти ограниченные цели устроена их картина мира и средства её корректировки.

Получается, что понимающему роботу нужны свои собственные неотъемлемые цели, то же выживание, т.е. самосохранение. Но не появятся ли тогда у него свои «три закона робототехники», где на первом месте будет он сам, а не человек? И будет ли возникшее у него понимание похоже на человеческое? Парадокс – стремясь получить человекоподобный интеллект, мы рискуем получить нечто принципиально «человеконеподобное». Нам это надо?

Заключение

На пути к сильному ИИ есть тупики, трудности и опасности. Как показано в разделе 1, МО – тупик. Это касается не только тех систем, которые обсуждаются в книге Ларсона, но и больших языковых моделей, появившихся позже. Более того, обсуждение понимания наводит на мысль, что тест Тьюринга не является тестом на способность мыслить: если, конечно, считать, что «мыслить» – это не только рассуждать, но и понимать, о чём ты рассуждаешь. Именно в формализации здравого смысла заключаются основные трудности на пути к сильному ИИ.

Невольно возникает вопрос: а может быть, любой цифровой путь к человекоподобному интеллекту – это тупик? Ведь мозг не вычисляет! Его механизмы хранения и обработки информации совсем не похожи на компьютерные механизмы. Уже неоднократно было замечено: то, что сложно человеку, просто компьютеру, и наоборот, то, что сложно компьютеру, просто человеку. Серьёзные прорывы в ИИ невозможны без прорывов в изучении информационных процессов мозга. При этом важно исследовать не только механизмы, управляющие этими процессами, но и их поразительную энергоэффективность.

Что касается опасностей, их можно разделить на две группы. Первая группа – это опасности, которые возникают из злоупотребления уже существующими возможностями ИИ. Вто-

рая группа опасностей носит экзистенциальный характер. После резкой критики (вместе с Ларсоном) мифов об ИИ, т.е. прогнозов Гуда-Курцвейла и их сторонников о грядущей сингулярности, размышления о понимании приводят к чему-то на первый взгляд похожему. Внешнее сходство действительно есть: в обоих случаях речь идёт о риске получить ИИ, не контролируемый человеком. Однако есть и принципиальная разница. Мифы об ИИ предполагают, что движение к сильному ИИ не контролируется уже сейчас, и что сильный ИИ неизбежно возникнет как бы сам собой в естественном ходе развития ИС; но современные тренды ИИ к сильному ИИ не ведут. Риски возможны на двух путях: самовоспроизводящиеся машины в смысле фон Неймана [9] и попытки наделить активные ИС пониманием в смысле, описанном в разделе 2, причём риски первого пути ничтожны и напоминают упоминавшуюся историю о печатающих обезьянах. Что же касается второго пути, то не исключено, что со временем придётся вводить ограничения и запреты, подобно тем, которые уже вводятся человечеством по отношению к экспериментам с человеческими эмбрионами и генетическим редактированием. Стоит прислушаться к предостережению Сета: «*Нам не стоит слепо и бездумно добиваться стандартной цели ИИ – воспроизвести, а затем превзойти человеческий интеллект. Мы создаём разумные инструменты, а не коллег. Если мы действительно внедрим в мир новые разновидности субъективного опыта, нам придётся иметь дело с нравственно-этическим кризисом беспрецедентных масштабов*» [28].

Список источников

[1] ***Erik J. Larson.*** The Myth of Artificial Intelligence. Why Computers Can't Think the Way We Do // The Belknap Press of Harvard University Press Cambridge, Massachusetts • London, England. 2021. 288 p.

[2] ***Shanahan Murray.*** The Technological Singularity. Cambridge, MA: MIT Press, 2015, 233 p.

[3] ***Good Irving John.*** Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers* 6 (1965) 6: 31–88.

[4] ***Глушков В.М.*** Интервью «Литературной газете». Литературная газета, 1975, №1.

[5] ***Vinge Vernor.*** The Coming Technological Singularity: How to Survive in the Post-Human Era // in Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, ed. G. A. Landis, NASA Publication CP-10129, 1993, 11–22.

[6] ***Kurzweil Ray.*** The Age of Intelligent Machines. The MIT Press. 1992. 565 p.

[7] ***Kurzweil Ray***. The Age of Spiritual Machines: When Computers Exceed Human Intelligence. Penguin Books, 2000. 404 p.

[8] ***Kurzweil Ray.*** The Singularity is Near: When Humans Transcend Biology. NY: Penguin Group, 2005. 652 p.

[9] ***Neumann John von.*** Theory of Self-Reproducing Automata. Edited and completed by Arthur W. Burks. University of Illinois Press, 1966. 403 p. Русский перевод: *Дж. Фон Нейман*. Теория самовоспроизводящихся автоматов. Закончено и отредактировано А. Берксом. М.: Мир, 1971. 382 с.

[10] ***Mange D., Stauffer A., Peparaolo L., Tempesti G.*** A Macroscopic View of Self-replication. Proceedings of the IEEE, 2004, **92** (12): 1929–1945.

[11] ***Russell Stuart.*** Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking, 2019. 352 p. Русский перевод: *Стюарт Рассел.* Совместимость. Как контролировать искусственный интеллект. М.: Альпина нон-фикшн, 2021. 438 с.

[12] ***Mitchell Melanie.*** Artificial Intelligence: A Guide for Thinking Humans/ New York: Farrar, Straus, and Giroux, 2019. 336 p. Русский перевод: *Митчелл Мелани*. Идиот или гений? Как работает и на что способен искусственный интеллект. М.: АСТ, 2022. 384 с.

[13] ***Levesque H.J., E. Davis, Morgenstern L.*** The Winograd Schema Challenge // Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, 2012. P.552-561.

[14] ***Davis E., Morgenstern L., Ortiz C.*** The Winograd Schema Challenge, https://cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html.

[15] ***Anderson Chris.*** The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired, June 23, 2008.

[16] ***Hill Sean.*** Simulating the Brain/ in Gary Marcus and Jeremy Freeman, eds., The Future of the Brain: Essays by the World's Leading Neuroscientists. Princeton, NJ: Princeton University Press, 2015, 123–124.

[17] ***Markram Henry.*** Seven Challenges for Neuroscience. Functional Neurology 28 (2013): 145–151.
[18] ***Wiener Norbert.*** Invention: The Care and Feeding of Ideas. Cambridge, MA: MIT Press, 1994. 159 p.
[19] ***Thompson N.C., Greenewald K., Lee K., Manso G.F..*** The Computational Limits of Deep Learning. arXiv:2007.05558v2 [cs.LG] 27 Jul 2022.
[20] ***Jonas E., Kording K.P.*** Could a Neuroscientist Understand a Microprocessor? / PLoS Comput Biol. 2017, 13(1): e1005268. DOI:10.1371/journal.pcbi.1005268.
[21] ***Осипов Г.С., Чудова Н.В., Панов А.И., Кузнецова Ю.М.*** Знаковая картина мира субъекта поведения. М.: Физматлит, 2018. 264 с.
[22] ***Uexküll J. von.*** Umwelt und Innenwelt der Tiere. Berlin: Verlag von Julius Springer, 1909. 276 p.
[23] ***Uexküll J. von.*** A Stroll through the Worlds of Animals and Men // Instinctive Behavior: The Development of a Modern Concept. N.Y.: International Universities Press, 1957. 328 p.
[24] ***Князева Е.Н.*** Понятие "Umwelt" Якоба фон Икскюля и его значимость для современной эпистемологии // *Вопросы философии*, 2015, № 5, 30-44.
[25] ***Lakoff G.*** Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press, 1987. 632 p. Русский перевод: *Лакофф Дж.* Женщины, огонь и опасные вещи: что категории языка говорят нам о мышлении. М.: Гнозис, 2011. 512 с.
[26] ***Pinker S.*** The Stuff of Thought: Language as a Window into Human Nature. NY: Viking, 2008. 512 p. Русский перевод: *Пинкер С.* Субстанция мышления: Язык как окно в человеческую природу. М.: Книжный дом «ЛИБРОКОМ», 2013. 557 с.
[27] ***Damasio A.R.*** Descartes' error: emotion, reason, and the human brain / Putnam Publishing, 1994. 312 p.
[28] ***Seth A.*** Being You: A New Science of Consciousness. Faber and Faber. 2021. 352 p. Русский перевод: *Анил Сет.* Быть собой. Новая теория сознания. М.: Альпина нон-фикшн, 2024. 400 с.
[29] ***Кузнецов О.П.*** О концептуальной семантике // *Искусственный интеллект и принятие решений*. 2014, №3, с.32-39.
[30] ***Кузнецов О.П.*** Когнитивная семантика и искусственный интеллект // *Искусственный интеллект и принятие решений*. 2012, № 4, с.32-42.
[31] ***Festinger Leon.*** A Theory of Cognitive Dissonance. Stanford University Press, 1962. 291 p. Русский перевод: *Фестингер Л.* Теория когнитивного диссонанса. Москва: Эксмо, 2018. 251 с.
[32] ***Kahneman D.*** Thinking, fast and slow. Farrar, Straus and Giroux, 2011. 499 p. Русский перевод: *Канеман Д.* Думай медленно … решай быстро. М.: АСТ, 2013. 653 с.
[33] ***Кузнецов О.П.*** Формальный подход к понятию «знание» и проблема моделирования различных типов знания // *Когнитивные исследования*. Сб. науч. тр. Вып. 2, М.: Институт психологии. 2008, с.265-275.
[34] ***Кузнецов О.П.*** Ограниченная рациональность и принятие решений // *Искусственный интеллект и принятие решений*. 2019, № 1, с.3-15.
[35] ***Sowa J.F.*** Conceptual Structures - Information Processing in Mind and Machines. Addison-Wesley Publ.Comp., 1984. 481 p.
[36] ***Richard J.F.*** Les activités mentales. Comprendre, raisonner, trouver des solutions/ Armand Colin, 1990. 446 p. Русский перевод: *Ж.Ф. Ришар*. Ментальная активность. Понимание, рассуждение, нахождение решений. М.: Институт психологии РАН, 1998.
[37] ***Minsky M.*** A Framework for Representing Knowledge / in. Winston P. (ed.), The Psychology of Computer Vision. N.Y., Mc Graw Hill, 1975, pp. 211-277. Русский перевод: М. Минский. Фреймы для представления знаний. – М.: Энергия, 1979.
[38] ***Rosch E.*** Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 1975. 104, P.192-233.
[39] ***Адамар Ж.*** Исследование психологии процесса изобретения в области математики. Пер. с фр. М. А. Шаталовой и О. П. Шаталова; Под ред. И. Б. Погребысского. М.: Сов. радио, 1970. 150 с.
[40] ***Карпов В.Э.*** Эмоции и темперамент роботов. Поведенческие аспекты // Известия РАН. Теория и системы управления. 2014. № 5. С.126–145.

Сведения об авторе

Кузнецов Олег Петрович 1936 г. рождения. Окончил МГУ им. М.В. Ломоносова философский факультет (1958), механико-математический факультет (1966). Доктор технических наук, профессор. Главный научный сотрудник Института проблем управления РАН. Автор более 170 публикаций, в том числе 4 монографий. SPIN-код: 4017-3236, AuthorID: 24. ORCID 0000-0002-5061-3855. *olpkuz@yandex.ru.*

Поступила в редакцию 14.09.2024, после рецензирования 3.10.2024. Принята к публикации 5.10.2024.

Scientific article DOI: 10.18287/2223-9537-2024-14-4-466-482

On machine learning, myths about General AI, and what understanding is

© 2024, O.P. Kuznetsov

Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia

Abstract

The first part of the article discusses the book *The Myth of Artificial Intelligence* by American scientist and entrepreneur E. Larson, which focuses on debunking some myths about artificial intelligence. These myths, which have persisted for over half a century, suggest that the emergence of human-like ("general") AI and eventually superintelligence is inevitable, occurring naturally as AI systems evolve. The book criticizes these myths in two ways: scientific and social. It is shown that machine learning does not lead to general AI, and the myth of AI makes human potential look weaker. The second part of the article considers the problem of understanding. The concept of cognitive semantics is proposed, based on the ideas of J. Lakoff, S. Pinker, A. Damasio and A. Seth. In particular, it is noted that: understanding is an interpretation in terms of a person's picture of the world; the picture of the world is constructed by our brain, and it is structured through the categorization of human experience; meanings (senses) are formed earlier than conceptual structures are formed; biological goals underlie meanings; not only the brain but also the body participates in cognitive processes, and understanding is associated with actions in the environment, knowledge of which is contained in the picture of the world. The article concludes by pointing out dead ends, difficulties and dangers on the path to general AI.

Keywords: *artificial intelligence, machine learning, induction, abduction, understanding, cognitive semantics, picture of the world.*

For citation: *Kuznetsov O.P.* On machine learning, myths about General AI, and what understanding is [In Russian]. *Ontology of designing*. 2024; 14(4): 466-482. DOI: 10.18287/2223-9537-2024-14-4-466-482.

Conflict of interest: The author declares no conflict of interest.

References

[1] ***Larson EJ.*** The Myth of Artificial Intelligence. Why Computers Can't Think the Way We Do // The Belknap Press of Harvard University Press Cambridge, Massachusetts • London, England 2021. 288 p.
[2] ***Murray S***. The Technological Singularity. Cambridge, MA: MIT Press, 2015, 233 p.
[3] ***Good IJ***. Speculations Concerning the First Ultraintelligent Machine. Advances in Computers 6 (1965) 6: 31–88.
[4] ***Glushkov VM.*** Interview with Literary newspaper [In Russian]. *Literaturnaya gazeta*, 1975, №1.
[5] ***Vernor V.*** The Coming Technological Singularity: How to Survive in the Post-Human Era // in Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, ed. G.A. Landis, NASA Publication CP-10129, 1993, 11–22.
[6] ***Kurzweil R.*** The Age of Intelligent Machines. The MIT Press. 1992. 565 p.
[7] ***Kurzwei. R.*** The Age of Spiritual Machines: When Computers Exceed Human Intelligence. Penguin Books, 2000. 404 p.
[8] ***Kurzweil R.*** The Singularity is Near: When Humans Transcend Biology. NY: Penguin Group, 2005. 652 p.
[9] ***Neumann JVN.*** Theory of Self-Reproducing Automata. Edited and completed by Arthur W. Burks. University of Illinois Press, 1966. 403 p.
[10] ***Mange D, Stauffer A, Peparaolo L, Tempesti G.*** A Macroscopic View of Self-replication. Proceedings of the IEEE, 2004, **92** (12): 1929–1945.
[11] ***Russell S.*** Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking, 2019. 352 p.
[12] ***Melanie M***. Artificial Intelligence: A Guide for Thinking Humans / NY: Farrar, Straus, and Giroux, 2019. 336 p.
[13] ***Levesque HJ, Davis E, Morgenstern L.*** The Winograd Schema Challenge // Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, 2012. P.552-561.
[14] ***Davis E, Morgenstern L, Ortiz C.*** The Winograd Schema Challenge, https://cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html.
[15] ***Anderson C.*** The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired, June 23, 2008.

[16] ***Hill S.*** Simulating the Brain/ in Gary Marcus and Jeremy Freeman, eds., The Future of the Brain: Essays by the World's Leading Neuroscientists. Princeton, NJ: Princeton University Press, 2015, 123–124.
[17] ***Markram H.*** Seven Challenges for Neuroscience. Functional Neurology 28 (2013): 145–151.
[18] ***Wiener N.*** Invention: The Care and Feeding of Ideas. Cambridge, MA: MIT Press, 1994. 159 p.
[19] ***Thompson NC, Greenewald K, Lee K, Manso GF***. The Computational Limits of Deep Learning. arXiv:2007.05558v2 [cs.LG] 27 Jul 2022.
[20] ***Jonas E, Kording KP.*** Could a Neuroscientist Understand a Microprocessor? *PLoS Comput Biol*. 2017, 13(1): e1005268. DOI:10.1371/journal.pcbi.1005268.
[21] ***Osipov GS, Chudova NV, Panov AI, Kuznetsova YuM***. The Symbolic Picture of the World of the Subject of Behavior. [In Russian]. Moscow: Fizmatlit, 2018. 264 p.
[22] ***Uexküll J. von.*** Umwelt und Innenwelt der Tiere. Berlin: Verlag von Julius Springer, 1909. 276 p.
[23] ***Uexküll J. von.*** A Stroll through the Worlds of Animals and Men // Instinctive Behavior: The Development of a Modern Concept. N.Y.: International Universities Press, 1957. 328 p.
[24] ***Knyazeva EN***. Jakob von Uexküll's Concept of "Umwelt" and its Relevance for Modern Epistemology [In Russian]. *Voprosy filosofii,* 2015; 5: 30-44.
[25] ***Lakoff G.*** Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press, 1987. 632 p.
[26] ***Pinker S.*** The Stuff of Thought: Language as a Window in to Human Nature. NY: Viking, 2008. 512 p.
[27] ***Damasio AR.*** Descartes' error: emotion, reason, and the human brain. Putnam Publishing, 1994. 312 p.
[28] ***Seth Al.*** Being You: A New Science of Consciousness. Faber and Faber. 2021. 352 p.
[29] ***Kuznetsov OP.*** Conceptual semantics [In Russian]. *Artificial Intelligence and Decision Making*. 2015; 42(5): 307-312.
[30] ***Kuznetsov OP.*** Cognitive semantics and artificial intelligence [In Russian]. *Artificial Intelligence and Decision Making*. 2013. T. 40, № 5. C. 269-276.
[31] ***Festinger L.*** A Theory of Cognitive Dissonance. Stanford University Press, 1962. 291 p.
[32] ***Kahneman D.*** Thinking, fast and slow / Farrar, Straus and Giroux, 2011. 499 p.
[33] ***Kuznetsov OP.*** Formal approach to the concept of "knowledge" and the problem of modeling different types of knowledge [In Russian]. *Cognitive Studies*. V. 2, Moscow: Institute of Psychology. 2008. P.265-275.
[34] ***Kuznetsov OP.*** Bounded rationality and decision making [In Russian]. *Artificial Intelligence and Decision Making*. 2019; 1: 3-15.
[35] ***Sowa JF.*** Conceptual Structures - Information Processing in Mind and Machines. Addison-Wesley Publ.Comp., 1984. 481 p.
[36] ***Richard JF.*** Les activités mentales. Comprendre, raisonner, trouver des solutions/ Armand Colin, 1990. 446 p.
[37] ***Minsky M***. A Framework for Representing Knowledge / in. Winston P. (ed.). The Psychology of Computer Vision. N.Y., Mc Graw Hill, 1975. P.211-277.
[38] ***Rosch E***. Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 1975; 104: 192-233.
[39] ***Hadamard J.*** Essai sur la psychologie de l'invention dans le domaine mathematique. Paris, 1959. 134 p.
[40] ***Karpov VE.*** Emotions and temperament of robots. Behavioural aspects [In Russian]. *Izvestiya RAS. Theory and control systems*. 2014; 5: 126–145.

About the author

Oleg Petrovich Kuznetsov (b.1936) graduated from the Lomonosov Moscow State University, Faculty of Philosophy (1958), Faculty of Mechanics and Mathematics (1966). He is a Doctor of Technical Sciences, a Professor, and a Chief Researcher at the Institute of Control Sciences of the Russian Academy of Sciences. He is the author of more than 170 publications, including 4 monographs. SPIN-код: 4017-3236, AuthorID: 24. ORCID 0000-0002-5061-3855. *olpkuz@yandex.ru*

Received September 14, 2024. Revised October 3, 2024. Accepted October 5, 2024.