

Содержание

1. Введение · 7
2. Краткая история превосходящего человека ИИ · 17
3. Интеллект (естественный) · 35
4. Интеллект (искусственный) · 55
5. Разновидности общего искусственного интеллекта · 71
6. Практическая разработка ОИИ · 85
7. Самообучение базовых моделей · 103
8. Открытое обучение в виртуальных мирах · 127
9. ОИИ и сознание · 145
10. Сверхинтеллект и взрывной рост интеллекта · 155
11. ОИИ и общество · 165
12. Заключение · 183

Благодарности · 198

Глоссарий · 199

Литература для дополнительного чтения · 205

1

Введение

ПОНЯТИЕ искусственного интеллекта (ИИ) многогранно. Это и давняя мечта человечества создать мыслящие машины, не уступающие человеку в способности рассуждать. Это и образы из массовой культуры: Терминатор, ВАЛЛ-И и Дейта из сериала «Звездный путь: следующее поколение». В научном мире ИИ — это предмет теоретических и экспериментальных исследований в университетских лабораториях. А в повседневной жизни — это уже существующие технологии, которые прокладывают маршруты, обрабатывают фотографии, управляют игровыми персонажами, ищут информацию в интернете и таргетируют рекламу на основе наших запросов.

В общественных дискуссиях часто происходит путаница между разными аспектами ИИ. Когда видишь заголовок «ИИ захватит мир», а потом смотришь на работу автодополнения в телефоне, становится очевидно, что такая несовершенная система едва ли способна на мировое господство. Или когда выбираешь автомобиль с ИИ-ассистентом вождения, может возникнуть необоснованное беспокойство, что система обретет собственный разум, как компьютер HAL 9000 из фильма «2001 год: Космическая одиссея». Существует явный диссонанс между статьями о превосходстве искусственного интеллекта над человеком и реальными экспериментами, где боты де-

монстрируют примитивное поведение, бесцельно двигаясь по кругу и натываясь на препятствия. Это вызывает сомнения в том, что авторы громких заголовков когда-либо сталкивались с реальной разработкой систем ИИ.

Один источник путаницы — это смешение реальных технологий с художественным вымыслом. Другой — смешение систем, выполняющих одну задачу, с системами, способными выполнять множество разных задач, даже таких, которым они не были обучены. Системы ИИ в фильмах «2001 год: Космическая одиссея», «Терминатор» и «ВАЛЛ-И» способны делать необычайно много вещей. Так, Терминатор может имитировать голоса только что встреченных людей, разрабатывать сложные планы, проводить расследования, управлять мотоциклом и сражаться с множеством полицейских, используя различное оружие. В то же время AlphaGo, система ИИ, поразившая мир в 2016 году победой над лучшими игроками в го, умеет только играть в го. Она не способна играть даже в шахматы без полного переобучения, не говоря уже об управлении мотоциклом или сражении с полицией.

Эта книга посвящена *общему искусственному интеллекту* (artificial general intelligence), или *ОИИ* (AGI), понятию, которое до сих пор не имеет единого определения. Несмотря на многочисленные попытки определить *ОИИ*, существующие формулировки часто расплывчаты и противоречивы. Билл Гейтс дает краткое определение: «программное обеспечение, способное обучаться любой задаче или предмету».¹ Джон Кармак, легендарный игровой технолог

1. Bill Gates, «The Age of AI Has Begun», *GatesNotes* (blog), March 21, 2023, <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun>.

и один из создателей Doom, предлагает иное видение: «форма ИИ, которая не просто имитирует человеческий интеллект, а понимает суть вещей и решает проблемы».² Исследователь ИИ Мюррей Шанахан определяет *ОИИ* как «искусственный интеллект, не ограниченный специфическими задачами, а способный осваивать такой же широкий спектр задач, как человек».³ Нетрудно найти системы, соответствующие одному определению, но не подходящие под другое. Существует еще множество определений — исследователь ИИ Бен Герцель проанализировал десятки из них в своей работе 2014 года, а дискуссия с тех пор стала только интенсивнее.⁴

В этой книге я буду использовать *ОИИ* для обозначения программного обеспечения, способного выполнять широкий спектр задач и решать разнообразные проблемы, существенно превосходя возможности современных систем ИИ. Это принципиально отличает его от *узкого ИИ* (*narrow AI*), который ограничивается выполнением одной или нескольких связанных между собой задач. Все или практически все существующие системы ИИ относятся именно к категории узкого ИИ.

Искусственный интеллект — еще один термин, не имеющий единого общепринятого определения. С 1960-х годов, когда этот термин появился, было

2. Glenn Hunter, «Exclusive Q&A: John Carmack's 'Different Path' to Artificial General Intelligence», Dallas Innovates, February 2, 2023, <https://dallasinnovates.com/exclusive-qa-john-carmacks-different-path-to-artificial-general-intelligence/>.

3. Murray Shanahan, *The Technological Singularity* (Cambridge, MA: MIT Press, 2015); Мюррей Шанахан, *Технологическая сингулярность* (Москва: Издательская группа Точка, 2017).

4. Ben Goertzel, «Artificial General Intelligence: Concept, State of the Art, and Future Prospects», *Journal of Artificial General Intelligence* 5, no. 1 (2014): 1.

Программа
для игры
в шахматы
имеет мало
общего с про-
граммой для
создания
реалистич-
ных изобра-
жений

предложено множество определений, нередко противоречащих друг другу. Это разнообразие частично объясняется множеством вычислительных подходов, развивавшихся в рамках искусственного интеллекта.⁵ Одни определения фокусируются на моделировании биологического или человеческого интеллекта, другие опираются на психологические концепции интеллекта, третьи концентрируются на конкретных вычислительных понятиях вроде символов или функций, а четвертые рассматривают типы задач, доступных для решения системами ИИ. В этой книге я буду использовать термин «искусственный интеллект» в двух значениях. Первое значение — стремление научить машины делать то, в чем люди пока их превосходят, особенно задачи, которые у людей требуют мыслительной деятельности. (Под «машинами» сейчас подразумеваются цифровые компьютеры.) Второе значение ИИ — совокупность алгоритмов, программного обеспечения и концепций, созданных исследователями и программистами для решения задач, которые, как считается, требуют интеллекта. Хотя книга рассматривает ИИ прежде всего как путь к созданию машинного интеллекта, я буду часто обращаться к отдельным уже существующим технологиям ИИ.

В этой книге вы не встретите, как и не должны встречать у компетентных авторов, выражение «искусственный интеллект» в значении некоего единого объекта или сущности. Такой вещи в мире сегодня просто не существует — есть только различные программные и аппаратные разработки, созданные для решения конкретных задач. Чтобы избежать недо-

5. Shane Legg and Marcus Hutter, «A Collection of Definitions of Intelligence», *Frontiers in Artificial Intelligence and Applications* 157 (2007): 17.

понимания, говоря о результатах исследований в области ИИ, я буду использовать термины «системы ИИ», «инструменты ИИ» и «алгоритмы ИИ». Важно понимать, что различные технологии, обозначаемые как «ИИ», часто принципиально различны: например, программа для игры в шахматы имеет мало общего с программой для создания реалистичных изображений.

Эта книга посвящена ОИИ, и поскольку таких систем пока не существует — даже самые продвинутые системы ИИ все еще довольно ограничены, — возникает вопрос, с чего же тогда начать. Логичной отправной точкой становится история, но не ИИ в целом, а история систем ИИ, превосходящих человеческие возможности. В следующей главе я рассмотрю шестидесятилетнюю историю ИИ через призму нескольких систем, превзошедших человека. Главный тезис главы состоит в том, что превосходство над человеком в отдельной области, каким бы впечатляющим и полезным оно ни было, не обязательно приближает нас к созданию общего интеллекта. Параллельно я постараюсь показать многообразие технических методов, используемых в различных системах ИИ.

Мы обсуждаем искусственный интеллект, но еще не определили само понятие «интеллект». В главе 3 речь пойдет об интеллекте неискусственного типа. Интеллект исследовали представители разных наук: философы, психологи, этологи (специалисты по поведению животных) и другие. Мы рассмотрим основные концепции в изучении человеческого и животного интеллекта и их отношение к общему искусственному интеллекту. В частности, мы обсудим фундаментальное разделение на текучий и кристаллизованный интеллект, помогающее понять различия между методами ИИ. Также мы рассмотрим

g-фактор — показатель, используемый для измерения общего интеллекта человека.

Кроме того, мы можем обратиться к теоретическим разработкам в области ИИ. В главе 4 я анализирую попытки сообщества исследователей ИИ определить понятие ОИИ. Здесь же рассматриваются математические аргументы, как поддерживающие, так и опровергающие возможность создания ОИИ. Среди них — парадоксальные теоремы о «бесплатном обеде», которые указывают на невозможность общего интеллекта, и концепция универсального интеллекта, предлагающая способ обойти это теоретическое ограничение.

Все эти теории об интеллекте и искусственном интеллекте могут показаться чем-то совсем абстрактным. Они не всегда помогают представить, как будет выглядеть ОИИ. В главе 5 мы рассмотрим различные версии того, каким может быть ОИИ, обращаясь к примерам из научной фантастики. Научная фантастика, вдохновлявшая целые поколения исследователей ИИ, помогает не только представить возможные варианты будущего с ИИ, но и понять различия между ними. Оставляя теоретические рассуждения и переходя к практическим исследованиям ИИ, я посвящаю следующие три главы реальным подходам к созданию ОИИ. В главе 6 я анализирую основные направления исследований ИИ и пытаюсь понять, способно ли хотя бы одно из них самостоятельно привести к созданию ОИИ. Забегая вперед: ответ отрицательный.

В следующих двух главах я рассматриваю наиболее перспективные, на мой взгляд, подходы к созданию ОИИ. Первый из них, описанный в главе 7, — базовые модели на основе самообучения (self-supervised learning). Последние годы принесли впечатляющие результаты в обучении крупных нейронных сетей

Как гипотетическая система ОИИ повлияет на занятость и наше чувство самооценности?

на колоссальных объемах данных из интернета. Получившиеся сети демонстрируют порой неожиданные возможности, справляясь даже с задачами, для которых их специально не обучали. GPT-4 от OpenAI, изначально обученная только для предсказания текста, способна выполнять переводы, делать обобщения, вести философские рассуждения, решать домашние задания и многое другое. Другая базовая модель, Stable Diffusion, может создавать разнообразные изображения по текстовым запросам, сжимать и ретушировать их и т.д. Однако нетрудно найти задачи, с которыми эти удивительные модели (пока) не справляются. Сработает ли такой подход в более крупном масштабе? Приведет ли простое увеличение размера сетей и объема данных для их обучения к созданию некой формы ОИИ?

В главе 8 я рассматриваю другой, комплементарный подход к ОИИ — открытое обучение в виртуальных средах. При таком подходе системы ИИ выступают как активные агенты в симулированных мирах, а не просто пассивно обучаются на созданных человеком данных. Идея основана на том, чтобы позволить ИИ развиваться эволюционным путем, подобно тому как развивался интеллект на Земле. Ученые давно изучают агентов, обучающихся действовать для максимизации вознаграждения, но несмотря на значительные успехи, такие агенты обычно вырабатывают ограниченные стратегии вместо универсальных решений. Особенность открытого обучения в том, что система ИИ сама формулирует свои цели и создает задачи, которые считает нужным решить. Это помогает избежать чрезмерной специализации. Хотя результаты обнадеживают, остается вопрос: смогут ли такие системы освоить навыки, полезные для человека?

Допустим, нам удастся создать систему искусственного интеллекта со значительной степенью общности. Будет ли у нее внутренний опыт существования? Будет ли у нее собственная точка зрения и самосознание? И если да, возникнут ли у нас этические обязательства перед такой системой? В главе 9 я рассматриваю сложную проблему сознания ОИИ. Хотя однозначных ответов в этой области нет, существуют заслуживающие внимания аргументы.

Опираясь на все это, мы можем обратиться к самой обсуждаемой идее в сфере ОИИ — той, что вызывает у людей одновременно восхищение и страх. Глава 10 рассматривает концепцию сверхинтеллекта (или взрывного роста интеллекта). Суть аргумента в том, что достаточно развитая система ИИ сможет совершенствоваться сама себя, становясь все умнее, что приведет к экспоненциальному росту интеллекта. По сравнению с таким интеллектом мы окажемся на уровне муравьев. Насколько обоснован этот аргумент? Согласуются ли его допущения с нашими знаниями об ИИ?

В главе 11 я анализирую социальные последствия появления ОИИ. Как гипотетическая система ОИИ повлияет на занятость и наше чувство самооценности? Ограничит ли она наши возможности управлять обществом? Анализ ОИИ в социальном контексте особенно важен, так как позволяет лучше осознать разнообразие человеческой деятельности и задач, с которыми должен будет справляться по-настоящему общий ИИ.

Наконец, в главе 12 я попробую обобщить содержание книги в емкий вывод. Подходя к концу книги, я вряд ли смогу удержаться от высказывания личного мнения. Острота суждений — умеренная.

Но это впереди. А пока вернемся к истокам цифровой эры.

Краткая история превосходящего человека ИИ

МАШИНЫ, превосходящие человеческий интеллект, уже существуют — они лучше большинства людей справляются с некоторыми задачами, для решения которых, как принято считать, требуется интеллект. Такие машины существуют уже долгое время, как минимум несколько десятилетий. Точная длительность этого периода зависит от того, какие именно задачи мы считаем требующими интеллекта. С годами это понимание менялось.

История искусственного интеллекта — попыток научить компьютеры выполнять действия, которые считались бы разумными, если бы их совершал человек, — восходит к 1940-м годам. Эта история не могла начаться раньше, так как до 1940-х не существовало цифровых компьютеров общего назначения. Практически сразу после создания первых таких компьютеров, которые по современным стандартам были громоздкими, медленными и крайне ограниченными, на них начали проводить эксперименты в области искусственного интеллекта. По сути, перспектива создания ИИ была одним из главных стимулов развития компьютеров. Этому способствовали как понимание потенциальной пользы от автоматизации различных видов мышления и обучения, так и извечное стремление человечества воссоздать интеллект в небиологической форме.

К ранним примерам разумных машин в художественных произведениях можно отнести искусственных слуг Гефеста из греческой мифологии и Голема — разумное существо из глины, созданное в XVI веке для защиты пражских евреев. После изобретения механических часов в XIV веке европейские часовщики начали создавать сложные автоматы, способные выполнять замысловатые движения. Их часто размещали в церквях, где они поражали воображение горожан и путешественников. Впрочем, эти автоматы были способны лишь на заранее запрограммированные движения, без какой-либо гибкости. Аналитическая машина Чарльза Бэббиджа представляла собой сложный программируемый калькулятор или ранний компьютер, но ее создатель успел разработать только чертежи — возможности механики XIX века не позволяли воплотить проект в реальность.

Появление цифровых технологий изменило ситуацию коренным образом. Благодаря скорости и надежности цифровых схем инженеры смогли создать вычислительные устройства с невиданными ранее возможностями. Первые программируемые цифровые компьютеры появились в 1940-х и 1950-х годах во многом благодаря потребностям Второй мировой войны в сложных вычислениях — для наведения артиллерии на движущиеся самолеты и обработки радиолокационных сигналов.¹ При всей их примитивности по современным меркам, эти машины творили чудеса. Прежде всего, они умели считать. До этого словом «компьютер» называли человека (как правило, женщину), который производил вычисления вручную или с помощью простейших механических

1. M. Mitchell Waldrop, *The Dream Machine* (Toronto: Stripe Press, 2018).

приспособлений. В крупных проектах, например при проектировании самолетов или артиллерии для военных целей, могли быть заняты сотни таких людей-«компьютеров». Появившиеся цифровые компьютеры выполняли те же расчеты в тысячи раз быстрее. Это была настоящая революция в вычислениях, но почему мы не считаем ее первым проявлением ИИ, превосходящего человека? Ведь вычисления — это работа нашего разума, и с 1950-х годов машины делают это лучше людей; даже самый простой калькулятор из дешевого магазина «думает» несопоставимо быстрее человека.

Некоторые люди, включая изобретателей, действительно называли эти новые вычислительные машины «электронными мозгами». Однако сегодня сложные вычисления сами по себе не считаются искусственным интеллектом. Одна из причин может быть в том, что они поддаются очевидной формализации: существуют четкие правила выполнения вычислений через разбиение их на последовательность простых элементарных операций. Возможно, именно эта очевидная механистичность процесса не позволяет нам считать карманные калькуляторы превосходящим человека ИИ, хотя человек и задействует мышление при вычислениях. Кроме того, сам термин «искусственный интеллект» появился уже после создания цифровых компьютеров, поэтому более ранние формы механического «мышления» могли не попасть в эту категорию.

Однако математика включает и другие когнитивные задачи, не только вычисления. Математики ведь не проводят все время за расчетами. Они развивают математику как науку. То есть доказывают теоремы и, что еще важнее, размышляют над тем, какие теоремы нужно доказывать и какими методами, что-

бы создавать полезные и красивые математические концепции. Если признать, что развитие математики требует интеллекта, то система, способная доказывать теоремы на уровне человека, должна считаться искусственным интеллектом, а если она делает это лучше человека, то ИИ, превосходящим человека.

В 1956 году Аллен Ньюэлл и Герберт Саймон решили проверить, способен ли компьютер доказывать математические теоремы. Цифровых компьютеров тогда было совсем немного, и каждый занимал целую комнату. Они сосредоточились на доказательствах в области оснований математики, выбрав знаменитую книгу начала XX века «Основания математики», которая стремилась свести математику к чистой логике.² Созданная ими программа Logic Theorist должна была заново вывести теоремы из «Оснований математики».³ По своей сути Logic Theorist представляла собой поисковый алгоритм, автоматически переписывающий доказательства согласно известным правилам преобразования уравнений, часть которых изучается еще в школе. Программа начинала с набора аксиом и последовательно преобразовывала их, пока не достигала намеченного доказательства. Существует колоссальное количество способов переписать эти аксиомы, но лишь немногие из них представляют ценность. Поэтому в Logic Theorist были заложены определенные правила и эвристики, имитирующие человеческий подход к поиску значимых способов

2. Alfred North Whitehead and Bertrand Russell, *Principia Mathematica* (Cambridge: Cambridge University Press, 1910); Альфред Н. Уайтхед и Бертран Рассел, *Основания математики: в 3 т.* Самара: Самарский университет, 2005.

3. Allen Newell and Herbert Simon, «The Logic Theory Machine — a Complex Information Processing System», *IRE Transactions on Information Theory* 2, no. 3 (1956): 61–79.

Конец ознакомительного фрагмента.
Приобрести книгу можно
в интернет-магазине
«Электронный универс»
e-Univers.ru