

Содержание

Предисловие	7
-------------------	---

Часть I

Разум как нейронная сеть

Глава 1	
Приглашение.....	15
Глава 2	
Как разум может возникнуть из мозга?.....	32
Глава 3	
Что делает нейронная сеть?.....	61

Часть II

Активация порождает мысль и действие

Глава 4	
Нейронная сеть как система памяти.....	101
Глава 5	
Контекст имеет значение.....	129
Глава 6	
То, что мы делаем.....	157

Часть III

Знание и обучение — все это в связях

Глава 7	
Как мы создаем (и теряем) смысл.....	187
Глава 8	
Эмерджентная мыслящая машина.....	229
Глава 9	
Когда мы учимся, мы изменяем связи.....	272

Часть IV

Расширение и применение нейросетевого подхода

Глава 10

Наши эмерджентные мысли.....311

Глава 11

Что означает концепция нейронных сетей для нас
и для искусственного интеллекта.....356

Благодарности.....385

Источники иллюстраций389

Предисловие

Наш мозг представляет собой огромную сеть клеток — *нейронов*, — которую оживляют волны электрической и химической активности, то нарастающие, то затухающие, то вновь набирающие силу. Наши восприятие, мысли, решения и действия — все то, что мы называем *разумом*, — рождаются из этих волн.

Как это происходит? Как разум может возникнуть из паттернов активности в мозге?

Для нас это один из самых древних и волнующих вопросов человечества. Он затрагивает саму нашу сущность и заставляет задуматься о месте человека во Вселенной. Он также связан с вопросом, существует ли искусственный разум и если да, то какова его природа. Мы назвали книгу «Рождение разума», чтобы предложить принципиально новый ответ на этот вопрос *как*.

Мы — ученые-практики, посвятившие свою жизнь изучению человеческого разума. Начиная наши исследования, поначалу независимо друг от друга, каждый из нас стремился выяснить, можно ли понять работу разума *механистически* — подобно тому как мы изучаем устройство самолета или развитие вирусного заболевания. Нам казалось, что существующие подходы к интересующим нас вопросам расплывчаты и слишком оторваны от конкретных фактов о работе мозга. Мы полагали, что опора на механистическое понимание процессов

в мозге поможет нам найти более точные ответы на вопросы о природе человека.

Физик Ричард Фейнман перед смертью оставил на доске такую запись: «Чего я не могу создать, того я не понимаю». Эта фраза точно отражает суть нашего подхода. Мы стремимся создавать системы, подобные мозгу, которые воспроизводят те феномены разума, которые мы хотим понять. Разумеется, воссоздать мозг во всех деталях невозможно. Поэтому мы строим модели, абстрагируясь от множества подробностей, чтобы продвигаться вперед.

Используемые нами модели — *модели нейронных сетей* — основаны на принципах работы обширных нейронных сетей мозга. Они позволяют исследовать, как человеческие способности могут возникать из нейронной активности. В этих моделях мы намеренно опускаем многие сложные подробности устройства мозга, чтобы сосредоточиться на базовых процессах, помогающих понять работу нашего разума. В этой книге мы расскажем о нейронных сетях, механизмы которых проливают свет на то, как люди воспринимают мир, принимают решения, формируют понятия и ставят цели.

Эти модели также помогают найти ответы на вопросы о человеческой природе, которые давно волнуют и нас, и многих других. Такие вопросы часто начинаются со слов *почему*, *что* и *откуда*. Почему мы порой не можем воплотить свои намерения в действия? Почему мы и окружающие находимся в плену предубеждений? Что позволяет нам иногда мгновенно схватывать истину, а в других случаях — совершенно не понимать очевидного? Откуда берется наша интуиция и почему она так часто нас подводит?

Поразительно, что нейронные сети, реализованные в виде компьютерных программ, стали фундаментом искусственного интеллекта. Модели, которые мы и другие исследователи

изначально создавали для понимания человеческого разума, легли в основу искусственных интеллектов. Поэтому изучение того, как нейронные сети воспроизводят наши мыслительные способности, проливает свет и на современные системы ИИ. В этой книге мы рассмотрим ключевые принципы нейросетевых моделей человеческого разума, на которых основаны системы ИИ, в некоторых аспектах достигающие наших когнитивных способностей, а порой и превосходящие их. Хотя мы сосредоточимся на тех идеях, которые считаем фундаментальными, мы также затронем некоторые стремительно развивающиеся инновации в системах ИИ середины 2020-х годов — особенно в двух заключительных главах. Эти инновации увлекательны и эффективны уже сейчас, но, вероятно, в ближайшие годы их ждет еще более стремительное развитие. Где возможно, мы описываем эти идеи с акцентом на их основополагающих принципах, а не на преходящих технических решениях. Наша цель — дать читателям фундамент, который останется актуальным, даже когда изменятся конкретные технологии.

Книга состоит из четырех частей. В первой части мы начнем с описания того, как система может обладать свойствами, которые отсутствуют у любой из ее составляющих. Этот феномен — *эмерджентность* — лежит в основе нейросетевой концепции разума, согласно которой разум возникает из взаимодействия простых вычислительных элементов, подобных клеткам мозга. Мозговые клетки сами по себе не способны мыслить, но их взаимодействие порождает мыслящую систему. Во второй части мы покажем, как нейросетевая концепция разума помогает объяснить широкий спектр человеческого поведения. Сначала мы рассмотрим нейросетевые модели памяти, включая ее несовершенство. Затем обратимся к тому, как мы, осмысляя окружающий мир, зависим от контекста, — в том

числе к тому, как ожидания формируют наши мысли. После этого мы исследуем процесс принятия решений — почему наш выбор иногда рационален, а иногда иррационален. В третьей части мы подробно расскажем, как нейронные сети — биологические и искусственные — обучаются на опыте. Мы опишем, как обучение формирует наши знания об объектах и их свойствах и как оно обеспечивает использование языка — особенно в больших языковых моделях. Наконец, в четвертой части мы расширим применение нейросетевого взгляда на разум. Мы покажем, как нейронные сети могут помочь в понимании таких феноменов, как формальное мышление, мотивированное поведение и сознание, — аспектов разума, которые пока не полностью охвачены нейросетевой теорией. В заключение обсудим некоторые следствия нейросетевого подхода — как для нас, людей, так и для наших машин.

Мы включили в книгу интерлюдии — воображаемые и вымышленные (если не указано иное) диалоги. Например, в одном из них Зигмунд Фрейд беседует с Адамом Смитом, в другом — редактор этой книги разговаривает с вымышленным посетителем бара в Нью-Йорке. Мы придумали эти разговоры, и все слова, вложенные в уста персонажей — исторических или вымышленных, — плод нашего воображения. Надеемся, эти интерлюдии оживят ваши размышления над поднятыми вопросами, так же как они оживили для нас процесс написания книги.

Мы создавали эту книгу для всех, кому интересен разум — человеческий и искусственный. Мы не предполагаем, что у читателя есть математическая подготовка. Кроме простого умножения и сложения, в книге нет ни одной формулы. Мы также не требуем предварительных знаний в области когнитивных, психологических, нейронных или компьютерных наук.

Предисловие

Наше понимание механизмов разума продолжает развиваться. Многие еще предстоит открыть — но то, что мы уже знаем, поражает воображение и имеет глубочайшие последствия.

Мы приглашаем вас в это путешествие. Возможно, оно изменит ваш взгляд на себя и на свое место во Вселенной.

Гаурав и Джей

ЧАСТЬ I

РАЗУМ КАК НЕЙРОННАЯ СЕТЬ

В первой части мы предлагаем вам рассмотреть идею о том, что разум можно продуктивно представить как результат взаимодействия клеток мозга, которые сами по себе не обладают способностями разума. Мы введем понятие *эмерджентности* — явления, при котором целое обладает свойствами, которые отсутствуют у его частей, — и покажем, каким образом модели нейронных сетей помогают понять, как возникает разум.

Глава 1

Приглашение

Когда одному из авторов этой книги, Гаураву, исполнилось 14 лет, родители подарили ему на день рождения сумму, эквивалентную примерно \$50. Этих денег хватило бы на джинсы клеш — абсолютный маст-хэв для любого подростка того времени — или на долгожданную школьную поездку с друзьями. Проблема заключалась в том, что он мог позволить себе что-то одно, а хотелось и то и другое. Нужно было принять решение. И вот тем вечером, собравшись с духом, он сделал выбор: он поедет в путешествие. Джинсы, в конце концов, могут подождать. Он был уверен, что поступает правильно.

Но наутро произошло нечто неожиданное: Гаурав проснулся с твердой уверенностью, что нужно выбрать джинсы. Варианты остались прежними, но его решение изменилось. Эти колебания, повторявшиеся несколько дней подряд, крайне озадачили его. В то время он только начинал осваивать компьютеры и представлял разум как своего рода компьютер, работающий по законам логики. Но какой компьютер выдает один ответ вечером, а другой — утром? Как могут меняться без видимой причины решения, основанные на логике? И если его мысли и предпочтения не результат работы логики и разума, то что же они такое?

Гаурав столкнулся с вечными вопросами человечества. Как рождаются наши мысли? Почему мы поступаем так, а не иначе? Можем ли мы доверять собственным мыслям? И в более общем смысле — что такое разум и как он работает?

Распространенные представления о разуме (и их недостатки)

Что такое разум? Можно сказать, что это сущность внутри нас, которая порождает наши мысли, восприятие, воспоминания, чувства, решения и действия. Но что он представляет собой на самом деле? Откуда берется? Рассмотрим вкратце несколько распространенных концепций и их недостатки.

Одна из концепций разума основана на религиозных традициях, многие из которых сходятся в том, что разум происходит от божественной материи или духовной субстанции, способной оживить человеческую плоть и превратить ее в личность. Нетрудно понять источник этой идеи: человеческая плоть кажется приземленной и бессмысленной. Как она может порождать интеллект? Несомненно, разум должен происходить от чего-то иного — чего-то неведомого, неземного. Чего-то, что связывает нас с бессмертным, священным и исполненным смысла. Чего-то божественного.

Проблема в том, что поиск корней разума в вечном, при всей возвышенности и даже красоте этой идеи, не объясняет, что такое разум и как возникают мысли. Такой подход превращает разум в невыразимую сущность, не поддающуюся дальнейшему познанию. Если наша цель — понять, как разум порождает мысли и все остальное, что мы ему приписываем, то останавливаться на этом нельзя.

Согласно второй концепции, разум представляет собой набор связанных друг с другом убеждений и желаний. Например, убеждение: «Люди с учеными степенями больше зарабатывают». Соответствующее желание: «Я хочу получить высокооплачиваемую работу». Логично предположить, что убеждения и желания взаимодействуют, порождая намерения и действия. Когда нас спрашивают, почему мы поступили определенным образом, мы можем сослаться на убеждения и желания, которые, по-видимому, повлияли на наше решение. Почему мы поступили в аспирантуру? Можно ответить: «Чтобы получить работу с хорошей зарплатой». Возможно, работа разума заключается во взаимодействии убеждений и желаний, формирующих цели, которые затем направляют наше поведение.

Недостаток модели «убеждения–желания» состоит в том, что она не объясняет происхождение этих убеждений и желаний. Как такие абстрактные явления, как убеждения и желания, возникают из физических процессов в мозге и как они порождают физические действия — движения, поступки, речь? Кроме того, модель не объясняет, почему люди часто действуют вопреки своим убеждениям и желаниям. Например, пациенты не принимают жизненно важные лекарства, а работники не открывают пенсионные счета, необходимые для финансовой стабильности. И это происходит несмотря на то, что эти люди верят в эффективность лекарств и хотят обеспечить себе достойную старость. Тем не менее они бездействуют.

Еще одна концепция представляет разум как программное обеспечение, которое получает данные из внешнего мира и применяет к ним набор правил, возможно, заложенных эволюцией. И действительно, в некоторых случаях работа разума напоминает применение правил: если у животного есть крылья и оно летает, мы, скорее всего, отнесем его к птицам; мы

выбираем блюдо в ресторанном меню, полагая, что оно будет лучше по сравнению с другими вариантами; мы образуем прошедшее время недавно возникшего глагола *fax* как *faxed* в соответствии с простым правилом: прошедшее время глагола в английском языке чаще всего образуется добавлением окончания *ed*.

Проблема концепции «разум как программа» в том, что она не продвигает нас далеко в понимании разума. Да, многие птицы летают, но мы узнаем птицу, даже если она не умеет летать. Да, мы часто выбираем в меню то, что нам нравится, но на наш выбор влияют факторы, не связанные с внутренней ценностью, — например, расположение блюда в верхней части страницы. Да, мы часто добавляем *ed* для образования прошедшего времени, но существует множество неправильных глаголов, где это правило не работает (например, *sleep* превращается в *slept*). Еще одно существенное ограничение этих моделей — они не привели к созданию работающих систем искусственного интеллекта. Подход, который мы представляем в нашей книге, оказался гораздо успешнее.

Согласно последней концепции, с которой мы часто сталкиваемся, различные аспекты разума зависят от отдельных «специалистов», каждый из которых располагается в своей специализированной области мозга. С этой точки зрения, мы двигаемся благодаря областям мозга, специализирующимся на движении; видим благодаря областям, отвечающим за зрение; чувствуем мотивацию благодаря областям, специализирующимся на мотивации; говорим благодаря языковым зонам — а в некоторых версиях этой теории мышление также осуществляется специализированными «мыслительными» областями.

Бесспорно, области мозга демонстрируют определенную степень специализации. Вопрос в том, откуда она

берется. Согласно одному подходу, который отстаивает философ Джерри Фодор в книге «Модулярность разума» (The Modularity of Mind), специализация возникает благодаря особым внутренним свойствам этих областей мозга, отобранным эволюцией для выполнения специфических вычислений. Хотя области мозга действительно в некоторой степени различаются по внутренней структуре, в нашей книге мы представляем иную точку зрения: специализация в основном является следствием различий между входящими и исходящими связями разных областей мозга. Например, часть мозга, называемая *зрительной корой*, играет важную роль в зрительном восприятии именно потому, что получает особенно мощный сигнал от глаз. Это означает, что изменение входящих сигналов должно изменить и функцию данной области мозга. И действительно, у людей, слепых от рождения, зрительная кора не получает визуальной информации и перепрофилируется для не визуальных задач — обработки слуховых или тактильных сигналов, которые также частично поступают в эту область.

Вместо представления о разуме как наборе жестко специализированных модулей такой подход предлагает видеть его как гибкую систему, формируемую опытом, обучением и требованиями среды. Это помогает объяснить существование того, что некоторые ученые называют «областью визуальной формы слова», — области мозга, которая специализируется на чтении написанных слов. Было бы странно утверждать, что эволюция создала эту область специально для чтения, ведь письменность была изобретена всего около 5000 лет назад. 5000 лет — слишком малый срок для эволюции, чтобы путем естественного отбора сформировать специализированный модуль чтения.

И все же у людей, умеющих читать, эта область мозга действительно специализируется на чтении. Почему? Чтение требует различения тончайших визуальных деталей — например,

различий между очень похожими буквами. Область визуальной формы слова получает мощные сигналы от нейронов зрительных областей мозга, обладающих максимальной чувствительностью к деталям, и потому вовлекается в процесс чтения. У людей, не умеющих читать, эта область специализируется на других задачах, также требующих различения мелких деталей, например на распознавании лиц. Такие факты подтверждают нашу точку зрения: специализация в мозге формируется опытом и зависит от входящих и исходящих связей областей мозга, а не от заранее заданных эволюцией функций.

Мы рассмотрели несколько концепций разума, каждая из которых обладает определенными недочетами. Теперь обратимся к концепции, которая открыла путь к гораздо более глубокому пониманию разума и сделала возможным развитие искусственного интеллекта: к представлению о разуме как о явлении, возникающем в нейронной сети.

Что такое нейронная сеть?

Осознав ограниченность привычных представлений о разуме, некоторые ученые и математики решили подойти к его пониманию совершенно по-новому. Они предположили, что, возможно, стоит начать с изучения продуктов деятельности разума, то есть мыслей и действий, путем отслеживания сигналов в сетях нейронов головного мозга. Мозг состоит из миллиардов клеток, которые называются *нейронами*. Нейрон — это базовый элемент, который обрабатывает и передает информацию в мозге и всей нервной системе.

На самом элементарном уровне функции нейронов вполне понятны. Они способны: 1) активироваться — то есть генерировать короткие электрические импульсы, называемые

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru