Потрясающе глубоко и при этом очень понятно.
Роберт Сапольски

Ясный и бесценный путеводитель по нейронным сетям.
Стивен Пинкер

Рождение разума

Как интеллект зарождается в людях и нейросетях

Гаурав Сури **Джей Макклелланд**

Джей Макклелланд
Гаурав Сури

Рождение разума: Как интеллект зарождается в людях и нейросетях

http://www.litres.ru/pages/biblio_book/?art=74101107
Рождение разума: Как интеллект зарождается в людях и нейросетях:
ISBN 9785006319349

Аннотация

Пытаясь решить проблему, мы обычно взвешиваем аргументы, полагаемся на интуицию, а затем делаем выбор. А что происходит в это время за кулисами – в нашем мозге?

Книга «Эмерджентный разум. Как рождается интеллект в людях и машинах» когнитивного психолога Джея Макклелланда и нейробиолога Гаурава Сури позволяет узнать, что такое человеческий разум с точки зрения нейросетевой модели и как он работает.

Что называют эмерджентностью? Как обучаются биологические и искусственные нейросети? Рациональны или иррациональны наши решения и от чего это зависит? Как ожидания формируют наши мысли?

В текст включены интерлюдии – воображаемые диалоги, которые оживляют повествование и вдохновляют на размышления. А еще можно потренироваться и построить собственную нейросеть.

Содержание

Гаурав Сури, Джей Макклелланд
Рождение разума: Как интеллект зарождается в людях и нейросетях

Знак информационной продукции (Федеральный закон № 436–ФЗ от 29.12.2010 г.)

12+

Переводчик: *Мария Кульнева*
Редактор: *Юлия Самуленок*
Главный редактор: *Сергей Турко*
Руководитель проекта: *Анна Деркач*
Арт-директор: *Юрий Буга*
Корректоры: *Елена Биткова, Евгений Яблоков*
Верстка: *Максим Поташкин*

* * *

Гаурав Сури Джей Макклелланд

Рождение разума

Как интеллект зарождается в людях и нейросетях

Перевод с английского

Альпина Паблишер

Москва, 2026

Нашим женам —
Ритике и Хайди

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 моз-

га. 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 находим-

ся 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 охваче-

ны 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.

Наше понимание механизмов разума продолжает развиваться. Многое 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, Гауравy, 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. Напри-

мер, 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 мно-

жество 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 невизуальных 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? Чте-

ние 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 отслежива-

ния 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 *потенциалами действия*; 2) передавать сигналы другим нейронам, с которыми соединены физически; 3) создавать новые связи с другими нейронами или изменять силу существующих.

Теперь попробуем представить себе сеть нейронов. Изображать настоящие нейроны довольно сложно, поэтому договоримся обозначать их кружками, которые будем называть *элементами* сети. Некоторые элементы связаны между собой, и эти связи мы можем показать стрелками. Элементы, которые активируются сигналами из внешнего мира, назовем *входными*; те, что передают сигналы во внешний мир, – *выходными*; а элементы без контакта с внешним миром – *скрытыми*. Мы изобразили нейронную сеть (рисунок 1.1).

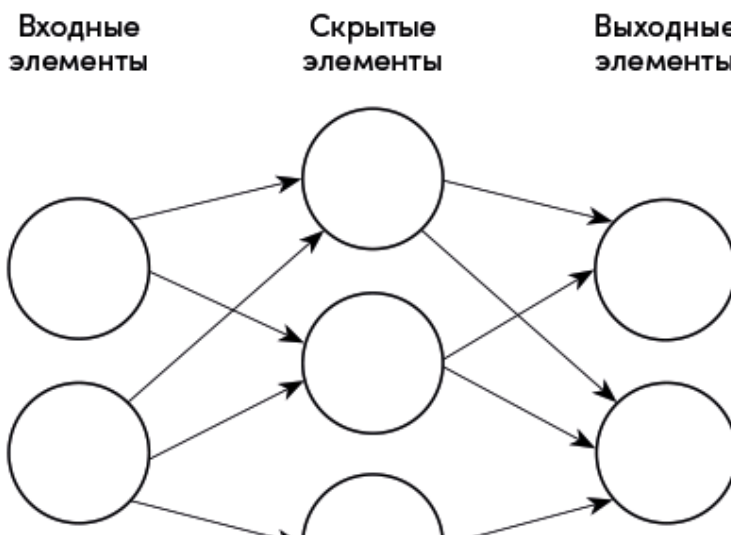

Рисунок 1.1. Примитивная модель нейронной сети

Как же такая нейронная сеть поможет нам понять разум? Чтобы показать возможности подобного подхода, представим автобус, несущийся на пешехода на переходе. Если рассматривать мозг пешехода как нейронную сеть, то сигналы, обнаруживающие большой движущийся объект, передаются через глаза к входным элементам. Те, в свою очередь, посылают сигналы скрытым элементам, которые активируют выходные элементы, заставляющие мышцы ног сократиться и

унести тело с опасного места. Стрелки здесь показывают направление передачи сигналов: например, входные элементы воздействуют на скрытые, поэтому между ними идут прямые стрелки.

Обратите внимание: это объяснение основывается на том, что мы знаем о возможностях нейронов. Получается, что наши нейроны – каждый из которых сам по себе не способен «думать» о том, чтобы увернуться от несущегося автобуса, – совместными усилиями осуществляют это действие, обмениваясь сигналами и активируя друг друга. Согласно теории нейронных сетей, все мысли и действия разума возникают благодаря взаимодействию нейронов. *Больше ничего нет.*

Разумеется, передача сигналов между нейронами в мозге устроена сложнее, чем в нашем упрощенном примере. Но можно достичь существенного прогресса, сосредоточившись на нескольких ключевых аспектах этого процесса. Во-первых, нейроны способны менять частоту генерации потенциалов действия – это можно представить как градуированную степень активации. Во-вторых, каждый нейрон способен принимать сигналы от множества других нейронов. По крайней мере в первом приближении можно сказать, что нейроны суммируют поступающие к ним сигналы активации и чем больше эта сумма, тем сильнее активируется сам нейрон. В-третьих, нейроны создают и изменяют силу связей в зависимости от опыта. Например, знакомство с новым человеком может сформировать связи между нейронами, кото-

рые представляют его лицо, и нейронами, которые представляют его имя. При повторных встречах эти связи укрепляются. Прочная связь между двумя нейронами позволяет им сильнее воздействовать друг на друга.

На схеме крайне сложно отобразить множество нейронов, активирующихся с разной интенсивностью и одновременно влияющих друг на друга через связи различной силы. Но одному из нас, Джею Макклелланду, помогла визуальная аналогия, которая позволила ему осмыслить одновременную работу элементов нейронной сети (рисунок 1.2).

Рисунок 1.2. Горный поток, падающий в природные резервуары, можно использовать как иллюстрацию некоторых аспектов одновременной работы нейронной сети

Джей был тогда молодым доцентом, который пытался понять разум не как систему, последовательно обрабатываю-

щую разрозненные идеи, а как непрерывный процесс, подверженный множеству воздействий различной интенсивности. Во время похода по Йосемитскому национальному парку, там, где другие люди видели горный поток, каскадами падающий из одного природного резервуара в другой, Джей вдруг увидел метафору непрерывных процессов разума. Он представил, что разные резервуары соответствуют разным мыслям. Количество воды в них соответствует их активации. Каждый водоем может получать воду из нескольких других, и уровень воды в нем определяется суммой всех этих струй. Одни бассейны соединены глубокими или широкими протоками, несущими мощные потоки, другие – едва заметными ручейками. Эти соединения напоминали связи различной силы, объединяющие мысли друг с другом. Он понял, что эта метафора позволяет думать о разуме как о нейронной сети, и вскоре начал строить модели, во многом опираясь на это интуитивное представление.

Метафора Джея, как и любая другая, несовершенна, но, помимо иллюстрации процессов в нейронной сети, ее механистическая природа указывает на заманчивую возможность: операции разума можно понять как физический процесс, подобный течению воды под уклон или повороту растений к свету, а не как проявление какой-то загадочной сущности. Не может ли это стать отправной точкой для разгадки всех тайн разума?

Возникновение разума в нейронной сети

Возможно, вы уже начинаете признавать, что нейронная сеть действительно открывает путь к пониманию разума. И все же вас может смущать мысль о том, что мышление возникает из взаимодействия нейронов, каждый из которых сам по себе мыслить не способен.

Центральная идея нашей книги: мышление *возникает* из компонентов, которые сами не могут мыслить.

Примеры помогут понять явление эмерджентности. Одно из самых впечатляющих проявлений эмерджентного поведения – согласованный полет птичьей стаи, обычно скворцов или ласточек. Стая движется как единое целое, пульсирует и колышется, создавая завораживающий воздушный балет (рисунок 1.3). Это явление называется мурмурацией, и, наблюдая за ним, легко поддаться ощущению, что стая – это единый организм с собственным разумом, который каким-то образом координирует движения всех своих частей.

Рисунок 1.3. Завораживающие сложные узоры птичьих стай не создаются дирижером. Они возникают из относи-

тельно простого поведения каждой отдельной птицы

Но никакого управляющего интеллекта нет. Каждая птица просто взаимодействует с несколькими ближайшими соседями в определенном радиусе. Каждая реагирует на изменения направления и скорости полета своих непосредственных соседей. Эти локальные взаимодействия создают эффект домино: малейшее изменение в полете одной птицы распространяется по стае, порождая грациозное скоординированное движение. Ни одна птица не знает об общем рисунке движения стаи – он возникает сам.

Это похоже на влажность воды. Ни одна молекула воды не обладает свойством влажности. Влажность возникает благодаря электрическому взаимодействию молекул воды друг с другом и с поверхностями, которых они касаются. Эмерджентность проявляется и в группах людей – толпа может быть способна на жестокость, на которую не способен ни один из ее участников в одиночку.

В этой книге мы покажем – без математики и технических подробностей, – как многое из того, что мы приписываем разуму (мысли, решения, эмоции), может возникать из взаимодействия нейронов. Мы еще многого не понимаем в этих процессах, но модели, созданные нами и другими исследователями, а также накопленные знания о мозге уже показывают, как наш опыт, мысли и действия могут рождаться в нейронной сети.

Путешествие, изменившее все

Любопытство Гаурава относительно природы его решений в конце концов привело его в аспирантуру Стэнфордского университета[1]. К тому времени он уже был знаком с нейронными сетями и эмерджентностью. Он интуитивно чувствовал потенциал этих идей, но еще не мог полностью связать их с волновавшими его вопросами.

И вот однажды Гаурав попал на семинар Джея Макклелланда, который к тому времени был уже маститым профессором.

Больше всего Гаурава поразила бескомпромиссная конкретность подхода Джея. Он не отмахивался от важных деталей, не ссылался на непостижимые теоретические концепции, не занимался спекуляциями и не делал удобных допущений. Вместо этого Джей выбирал явление, которое хотел понять, формулировал набор четких исходных принципов и предположений, а затем прозрачно применял эти принципы для вычисления активаций нейронной сети при заданных входных сигналах. Гаурав пришел к Джею, чтобы лучше понять его подход.

Первое явление, которое Джей продемонстрировал Гаураву, касалось извлечения информации о группе людей. Это была разработанная Джеем нейронная сеть, которую мы опи-

[1] Организация, деятельность которой признана нежелательной в РФ.

шем в главе 4. Гаурав тщательно проследил активации в сети и с восторгом обнаружил, что нейронная сеть способна делать логические выводы. Эта способность не была заложена в сеть явно – она возникла из взаимодействия элементов сети. Это стало поворотным моментом. Он помнит, как сидел за столом и сказал себе: «Я уже никогда не буду думать о разуме так, как думал раньше».

Цель нашей книги – дать читателям возможность пережить такое же откровение.

Для достижения этой цели мы отобрали результаты экспериментов с людьми, животными и искусственными нейронными сетями, которые проливают свет на важнейшие аспекты человеческого разума. Эти примеры показывают, как мы воспринимаем мир, формируем понятия, принимаем решения, испытываем эмоции, проявляем самоконтроль, используем и понимаем язык и даже как мы мыслим и рассуждаем. Каждое явление, которое мы стремимся объяснить, подкреплено обширными экспериментальными данными, позволяющими оценить объяснительную силу разрабатываемых нами нейросетевых моделей.

Дополнительный бонус: понимание искусственного интеллекта

Для нас стремление понять собственный разум – возвышенная цель сама по себе, которая не требует дополнитель-

ных стимулов помимо внутренней ценности самопознания. Но оказалось, что изучение разума через нейронные сети дает полезный и неожиданный бонус: понимание современных систем искусственного интеллекта, *которые основаны на тех же принципах нейронных сетей,* что изначально разрабатывались для понимания человеческого разума.

Современные системы ИИ намного превзошли ранние попытки создать машины, способные распознавать изображенные объекты. Они могут переводить с одного языка на другой, улавливая тончайшие смысловые оттенки, превосходят человека в глубоко интуитивных играх вроде го и ведут естественную, похожую на человеческую беседу на практически неограниченный круг тем. Этот список впечатляющих достижений наверняка будет расти в ближайшие годы.

На волне восторгов по поводу этих прорывов потерялось важное понимание: системы машинного обучения, демонстрирующие такие способности, – это, по сути, нейронные сети, во многом похожие на модели, которые мы и другие исследователи используем для изучения ключевых аспектов человеческого разума уже более 50 лет. Их интеллект основан на заимствованной у мозга концепции: совокупности нейроноподобных элементов, которые возбуждают и тормозят друг друга через связи. Несмотря на десятилетия усилий, предыдущие версии ИИ, опиравшиеся на символы, обозначающие объекты и отношения в соответствии с системами правил, не достигли уровня интеллекта, который теперь во-

площен в нейросетевом ИИ.

Конец ознакомительного фрагмента.