ДЖУДА ПЕРЛ
ДАНА МАККЕНЗИ

ДУМАЙ «ПОЧЕМУ?»

ПРИЧИНА И СЛЕДСТВИЕ КАК КЛЮЧ К МЫШЛЕНИЮ

ЗАДУМЫВАЛИСЬ ЛИ ВЫ КОГДА-НИБУДЬ О ЗАГАДКАХ КОРРЕЛЯЦИЙ И ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ? ЭТА ЗАМЕЧАТЕЛЬНАЯ КНИГА ОТВЕТИТ НА ВАШИ ВОПРОСЫ. КРОМЕ ТОГО, ЕЕ ВЕСЕЛО ЧИТАТЬ!
ДАНИЭЛЬ КАНЕМАН

Джудиа Перл
Дана Маккензи

Думай «почему?». Причина и следствие как ключ к мышлению

Серия «Власть и успех»

Текст предоставлен издательством
http://www.litres.ru/pages/biblio_book/?art=68683656
Думай «почему?». Причина и следствие как ключ к мышлению: АСТ; Москва; 2022
ISBN 978-5-17-123140-8, 978-5-17-146449-3

Аннотация

Удостоенный премии Алана Тьюринга 2011 года по информатике, ученый и статистик показывает, как понимание причинно-следственных связей произвело революцию в науке и совершило прорыв в работе над искусственным интеллектом.

«Корреляция не является причинно-следственной связью» – эта мантра, скандируемая учеными более века, привела к условному запрету на разговоры о причинно-следственных связях. Сегодня это табу отменено. Причинная революция, открытая Джудией Перлом и его коллегами, пережила столетие

путаницы и поставила каузальность – изучение причин и следствий – на твердую научную основу.

Работа Перла позволяет нам не только узнать, является ли одно причиной другого, она позволяет исследовать реальность, которая уже существует, и реальности, которые могли бы существовать. Она демонстрирует суть человеческой мысли и дает ключ к искусственному интеллекту.

В формате PDF A4 сохранен издательский макет книги.

Содержание

Дана Маккензи, Джудиа Перл
Думай «почему?». Причина и следствие как ключ к мышлению

Judea Pearl and Dana Mackenzie
The Book of Why: the New Science of Causes and Effect

Предисловие

Почти два десятилетия назад, работая над предисловием к книге «Причинность» (2000), я сделал довольно смелое замечание, после которого друзья посоветовали мне умерить пыл. Я написал: «Причинность пережила важнейшую трансформацию – от понятия, овеянного тайной, до математического объекта с хорошо определенным смыслом и хорошо обоснованной логикой. Парадоксы и противоречия были разрешены, туманные понятия были истолкованы, а связанные с причинностью практические задачи, которые долго считались или метафизическими, или нерешаемыми, теперь могут быть разрешены при помощи элементарной математики. Проще говоря, причинность была математизирована».

Перечитывая этот отрывок сегодня, я чувствую, что был весьма близорук. Явление, описанное мной как «трансформация», оказалось «революцией», которая изменила мышление ученых в самых разных науках. Многие сегодня называют это Революцией Причинности, и волнение, которое она вызвала в кругах исследователей, сейчас распространяется на образование и практическую сферу.

У этой книги тройная задача: во-первых, описать для вас нематематическим языком интеллектуальную суть Революции Причинности и показать, как она влияет на нашу жизнь и на будущее; во-вторых, рассказать о героических путеше-

ствиях, как успешных, так и неудачных, в которые отправились некоторые ученые, столкнувшись с важнейшими вопросами, касающимися причинно-следственных связей.

Наконец, возвращая Революцию Причинности к ее истокам в сфере искусственного интеллекта (ИИ), я ставлю целью показать вам, как можно создать роботов, способных общаться на нашем родном языке – языке причины и следствия. Это новое поколение роботов должно объяснить нам, почему случились определенные события, почему они откликнулись определенным образом и почему природа действует так, а не иначе. Более амбициозная цель – узнать от них, как устроены мы сами: почему наш ум срабатывает именно так и что значит думать рационально о причине и следствии, вере и сожалении, намерении и ответственности.

Когда я записываю уравнения, у меня есть очень четкое представление о том, кто мои читатели. Но если я пишу для широкой публики, его нет, и это для меня совершенно новое приключение. Странно, но такой новый опыт стал одним из самых плодотворных образовательных усилий в моей жизни. Необходимость выражать идеи на вашем языке, думать о вашем опыте, ваших вопросах и ваших реакциях обострила мое понимание причинности больше, чем все уравнения, которые я написал до того, как создал эту книгу.

За это я буду вечно благодарен. И надеюсь, что вам так же, как и мне, не терпится увидеть результаты.

Джудиа Перл, Лос-Анджелес, октябрь 2017 года

Введение: Ум важнее данных

Любая развитая наука смогла развиться благодаря собственным символам.
Огастес де Морган, 1864

Эта книга рассказывает историю науки, которая повлияла на то, как мы отличаем факты от вымысла, и осталась при этом вне поля зрения широкой публики. Новая наука уже определяет важнейшие аспекты нашей жизни и потенциально может повлиять на многое другое: от разработки новых лекарств до управления экономическим курсом, от образования и робототехники до контроля над оборотом оружия и глобальным потеплением. Примечательно, что, несмотря на разнообразие и явную несоизмеримость этих областей, новая наука собирает их все в рамках единой структуры, которой практически не существовало два десятилетия назад.

У нее нет красивого названия – я называю ее просто причинным анализом, как и многие коллеги. Не особо высокотехнологичный термин. Идеальная технология, которую пытается моделировать причинный анализ, есть у нас в голове. Десятки тысяч лет назад люди начали понимать, что одни вещи приводят к другим вещам и что, регулируя первое, можно повлиять на второе. Ни один биологический вид, кроме нашего, не осознает этого – по крайней мере, до такой сте-

пени. Это открытие породило организованные общества, потом города и страны и наконец-то цивилизацию, основанную на науке и технике, которая есть у нас сегодня. И все потому, что мы задали простой вопрос: почему? Причинный анализ относится к этому вопросу очень серьезно. Он исходит из предпосылки о том, что человеческий мозг – самый продвинутый инструмент из когда-либо созданных для работы с причинами и следствиями. Мозг хранит невероятный объем знаний о причинности, и, поддержав его данными, можно использовать этот орган для ответа на самые насущные вопросы нашего времени. Более того, как только мы действительно поймем логику, стоящую за рассуждениями о причинах, мы будем способны имитировать ее в современных компьютерах и создать «искусственного ученого». Этот умный робот откроет еще неизвестные феномены, найдет объяснения для неразрешенных научных дилемм, разработает новые эксперименты и будет постоянно извлекать новые знания о причинах явлений из окружающей среды.

Но прежде, чем мы начнем размышлять о подобных футуристических достижениях, важно понять достижения, к которым уже привел нас причинный анализ. Мы исследуем, как он преобразил мышление ученых почти во всех дисциплинах, основанных на работе с данными и как это вскоре изменит нашу жизнь. Новая наука занимается довольно однозначными на первый взгляд вопросами вроде таких:

• Насколько эффективно данное лечение для предотвра-

щения болезни?

• Что вызвало рост продаж – новый закон о налогообложении или наша рекламная кампания?

• Как ожирение влияет на траты на медицинское обслуживание?

• Могут ли данные о найме сотрудников служить доказательством последовательной дискриминации по половому признаку?

• Я собираюсь уволиться. Стоит ли это делать?

Во всех этих вопросах видна озабоченность причинно-следственными отношениями, которую можно узнать по таким словам, как «предотвращения», «вызвало», «влияет», «последовательной» и «стоит ли». Эти слова часто встречаются в повседневном языке, и наше общество постоянно требует ответы на эти вопросы. Но до недавнего времени наука не давала нам средств, чтобы даже выразить их, не говоря уже о том, чтобы на них ответить.

Наука о причинном анализе оставила это пренебрежение со стороны ученых в прошлом, и в этом состоит ее важнейшее достижение на благо человечество. Новая наука породила простой математический язык, чтобы выражать каузальные отношения – и те, о которых мы знаем, и те, о которых хотели бы узнать. Возможность выразить эту информацию в математической форме открыла изобилие мощных, основанных на твердых принципах методов, которые позволяют

сочетать наше знание с данными и отвечать на каузальные вопросы вроде пяти, приведенных выше.

Мне повезло участвовать в развитии этой научной дисциплины в течение последней четверти века. Я наблюдал, как она оформляется в студенческих аудиториях и исследовательских лабораториях, и видел, как ее прорывы сотрясают угрюмые научные конференции вдали от софитов общественного внимания. Сейчас, когда мы вступаем в эру сильного искусственного интеллекта, многие славят бесконечные возможности, которые открывают большие массивы данных и технологии глубинного обучения. Я же нахожу своевременной и волнующей возможность представить читателю смелые пути, которыми идет новая наука, и рассказать, как она влияет на науку о данных и какими разнообразными способами изменит нашу жизнь в XXI веке.

Вероятно, когда вы слышите, что я называю эти достижения новой наукой, у вас появляется скепсис. Вы можете даже спросить: почему она не появилась давным-давно? Например, когда Вергилий провозгласил: «Счастлив тот, кто смог понять причины вещей» (29 год до н. э.). Или когда основатели современной статистики Фрэнсис Гальтон и Карл Пирсон впервые открыли, что данные о населении могут пролить свет на научные вопросы. Кстати, за их досадной неспособностью учесть причинность в этот ключевой момент стоит долгая история, которую мы рассмотрим в исторических разделах этой книги. Однако самым серьезным препятстви-

ем, с моей точки зрения, было фундаментальное расхождение между языком, на котором мы задаем вопросы о причинности, и традиционным языком, которым описываем научные теории.

Чтобы оценить глубину этого расхождения, представьте трудности, с которыми столкнется ученый, пытаясь объяснить некоторые очевидные причинные отношения, скажем, что барометр, показывающий B, считывает давление P. Это отношение легко записать уравнением $B = kP$, где k – некий коэффициент пропорциональности. Правила алгебры теперь позволяют нам переписать это уравнение в самых разных формах, скажем $P = B/k$, $k = B/P$ или $B - kP = 0$. Все они означают одно и то же: если мы знаем любые две из трех величин, третья определена. Ни одна из букв k, B или P не имеет преимуществ перед остальными с математической точки зрения. Но как же выразить наше сильное убеждение в том, что давление заставляет показания барометра измениться, а не наоборот? А если мы не способны выразить даже это, как же сформулировать другие наши убеждения о причинно-следственных отношениях, у которых нет математических формул? Например, о том, что от кукареканья петуха солнце не встает?

Мои преподаватели в университете не могли этого сделать, но никогда не жаловались. Я готов поспорить, что ваши тоже. И сейчас мы понимаем почему: им никогда не показывали математический язык причинности и никогда не рас-

сказывали о его пользе. Более того, это обвинительный приговор науке, которая в течение стольких поколений игнорировала необходимость подобного языка. Все знают, что если щелкнуть выключателем, то зажжется свет, и что в жаркий и душный день в местном кафе-мороженом поднимутся продажи. Почему же ученые до сих пор не выразили такие очевидные факты в формулах, как это было сделано с базовыми законами оптики, механики или геометрии? Почему они допустили, чтобы эти факты чахли, ограниченные голой интуицией и лишенные математических инструментов, которые позволили другим наукам зреть и процветать?

Отчасти ответ в том, что научные инструменты развиваются, дабы удовлетворять научные потребности. Именно потому, что мы так хорошо управляемся с вопросами о выключателях, мороженом и барометрах, наша потребность в особых математических инструментах, чтобы их решать, была неочевидной. Но по мере того, как научное любопытство увеличилось и мы начали задавать вопросы о причинности в сложных юридических, деловых, медицинских и политических ситуациях, оказалось, что у нас не хватает инструментов и принципов, которые должна предоставить зрелая наука.

Запоздалое пробуждение такого рода нередко встречается в науке. Например, вплоть до середины XVII века люди вполне удовлетворялись своей способностью справляться с неопределенностью в повседневной жизни – от перехо-

да улицы до риска подраться. Только когда азартные игроки изобрели изощренные игры, порой тщательно нацеленные на то, чтобы вынудить других сделать неверный выбор, математики Блез Паскаль (1654), Пьер Ферма (1654) и Христиан Гюйгенс (1657) посчитали необходимым развить то, что сегодня мы называем теорией вероятностей. Подобным образом лишь тогда, когда страховым организациям потребовалось точно рассчитать пожизненную ренту, такие математики, как Эдмунд Галлей (1693) и Абрахам де Муавр (1725), использовали данные о смертности, чтобы вычислить ожидаемую продолжительность жизни. Аналогично потребности астрономов в точном предсказании движения небесных тел подтолкнули Якоба Бернулли, Пьера Симона Лапласа и Карла Фридриха Гаусса разработать теорию ошибок, которая помогает выделить сигналы из шума. Все эти методы – предшественники сегодняшней статистики.

Удивительно, но потребность в теории причинности начала оформляться в то же время, когда появилась статистика. Более того, современная статистика родилась из вопросов о причинах, которые Гальтон и Пирсон задавали применительно к наследственности, и из их изобретательных попыток на них ответить, используя данные о нескольких поколениях. К сожалению, попытка не удалась, и вместо того, чтобы остановиться и спросить почему, они объявили эти вопросы недоступными для изучения и занялись развитием процветающей, свободной от причинности области под на-

званием «Статистика».

Это был важнейший момент в истории науки. Возможность решать вопросы причинности на ее собственном языке почти воплотилась, однако ее растратили напрасно. В последующие годы эти вопросы были объявлены ненаучными и отправлены в подполье. Несмотря на героические усилия генетика Сьюалла Райта (1889–1988), вокабуляр причинности был буквально запрещен больше чем на 50 лет. А запрещая речь, вы запрещаете мысль и душите принципы, методы и инструменты.

Читателям этой книги не надо быть учеными, чтобы увидеть данный запрет своими глазами. Осваивая курс «Введение в статистику», каждый студент учится повторять: «Корреляция не означает причинно-следственную связь». И этому есть хорошее объяснение! Кукареку петуха тесно коррелирует с рассветом, но не является его причиной.

К сожалению, в статистике это здравое наблюдение стало фетишем. Оно сообщает нам, что корреляция не означает причинно-следственную связь, но не говорит нам, что такое эта причинно-следственная связь. Попытки найти раздел «Причина» в учебниках по статистике обречены на неудачу. Студентом не разрешается говорить, что X причина Y, – только что X и Y «связаны» или «ассоциируются».

Из-за этого запрета математические инструменты для работы с вопросами причинности были признаны излишними, и статистика сосредоточилась исключительно на обобщении

данных, а не на их интерпретации. Блестящим исключением стал путевой анализ, изобретенный генетиком Сьюаллом Райтом в 1920-е годы – прямой предок методов, которые мы рассмотрим в этой книге. Однако путевой анализ не получил должной оценки в статистике и сопряженных сообществах и десятилетиями пребывал в состоянии эмбриона. То, что должно было стать первым шагом по направлению к причинному анализу, оставалось единственным шагом до 1980-х годов. Остальная статистика, а также многие дисциплины, которые на нее ориентировались, так и жили в эпоху этого «сухого закона», ошибочно полагая, что ответы на все научные вопросы кроются в данных и должны быть открыты с помощью умных способов их интерпретировать.

Эта ориентация на данные до сих пор преследует нас. Мы живем в эпоху, когда большие данные считаются потенциальным решением для всех проблем. Курсы по теории и методам анализа данных в изобилии преподаются в наших университетах, а компании, участвующие в «экономике данных», готовы платить хорошие деньги специалистам в этих вопросах. Но я надеюсь убедить вас этой книгой, что данные – вещь крайне тупая. Они могут рассказать вам, что люди, которые приняли лекарство, восстановились быстрее, чем те, кто его не принимал, но не могут рассказать почему. Может, те, кто принял лекарство, сделали так, поскольку были в состоянии позволить это себе, но восстановились бы столь же быстро и без него.

Снова и снова в науке и бизнесе мы наблюдаем ситуации, в которых одних данных недостаточно. Большинство энтузиастов, работающих со значительными массивами данных, осознавая порой эти ограничения, продолжают ориентироваться на искусственный интеллект, обрабатывающий данные, как будто альтернатива все еще под запретом.

Как я говорил выше, за последние 30 лет ситуация радикально изменилась. Сегодня, благодаря тщательно созданным причинным моделям, современные ученые могут обратиться к проблемам, которые когда-то сочли бы нерешаемыми или даже не подходящими для научного изучения. Например, всего 100 лет назад вопрос о том, вредит ли здоровью курение сигарет, был бы признан ненаучным. Одно упоминание слов «причина» и «следствие» вызвало бы лавину возражений в любом авторитетном журнале о статистике.

Еще 20 лет назад задать статистику вопрос вроде «Это аспирин помог мне от головной боли?» было все равно, что спросить, верит ли он в магию вуду. Как выразился мой почтенный коллега, это была бы «скорее тема для светской беседы, а не научный запрос». Но сегодня эпидемиологи, обществоведы, специалисты по компьютерным наукам и, по крайней мере, некоторые просвещенные экономисты и статистики регулярно ставят такие вопросы и отвечают на них с математической точностью. Для меня эти перемены равнозначны революции. Я осмеливаюсь называть их Революцией Причинности, научной встряской, которая позволяет прини-

мать, а не отрицать наш врожденный когнитивный дар понимать причины и следствия.

Революция Причинности произошла не в вакууме; за ней стоит математический секрет, который лучше всего можно описать как численные методы причинности; они отвечают на самые сложные вопросы, когда-либо заданные о причинно-следственных отношениях. Я открываю эти методы с большим волнением – не только потому, что бурная история их появления весьма интригует, но и в большей степени потому, что, по моим ожиданием, в будущем их потенциал раскроют, опередив самые смелые мечты, и… вероятно, это сделает один из читателей настоящей книги.

Вычислительные методы причинности включают два языка: диаграммы причинности, которые выражают то, что мы знаем, и символический язык, напоминающий алгебру, который выражает то, что мы хотим узнать. Диаграммы причинности – простые рисунки из точек со стрелками, которые обобщают существующее научное знание. Точки символизируют интересующие нас факторы под названием «переменные», а стрелки – известные или подразумеваемые причинные отношения между ними, означающие, к каким переменным «прислушивается» та или иная переменная. Такие диаграммы невероятно легко рисовать, понимать и использовать, и читатели обнаружат их в изобилии на страницах этой книги. Если вы сможете найти дорогу по карте улиц с односторонним движением, то поймете диаграммы причин-

ности и ответите на вопросы, относящиеся к тому же типу, что и заданные в начале этого вступления.

Диаграммы причинности, которые я предпочитаю использовать в этой книге и выбираю в качестве основного инструмента в последние 35 лет, не единственная модель причинности. Некоторые ученые (например, специалисты по эконометрике) любят работать с математическими уравнениями, другие (скажем, закоренелые статистики) предпочитают список допущений, которые предположительно обобщают структуру диаграммы. Независимо от языка, модель должна описывать, пусть и качественно, процесс, который порождает данные, – другими словами, причинно-следственные силы действуют в среде и формируют порождаемые данные.

Бок о бок с этим диаграммным «языком знания» существует символический «язык запросов», на котором мы выражаем вопросы, нуждающиеся в ответах. Так, если нас интересует эффект лекарства (*D – drug*) на продолжительность жизни *(L – lifespan*), то наш запрос можно символически записать так: *P* (*L* | *do (D)*). Иначе говоря, какова вероятность (*P – probability*) того, что типичный пациент проживет *L* лет, если его заставят принимать это лекарство? Вопрос описывает то, что эпидемиологи назвали бы интервенцией или лечением, и соответствует тому, что мы измеряем во время клинического исследования. Во многих случаях мы также захотим сравнить *P* (*L* | *do* (*D*)) и *P* (*L* | *do* (*не-D*)); последнее в данном случае описывает пациентов, которые не получи-

ли лечения, так называемую контрольную группу. Оператор *do* означает, что мы имеем дело с интервенцией, а не с пассивным наблюдением. В классической статистике нет ничего даже напоминающего этот оператор.

Мы должны применить оператор интервенции $do\,(D)$, чтобы убедиться: наблюдаемое изменение в продолжительности жизни L объясняется самим лекарством и не объединено с другими факторами, которые могут укорачивать или удлинять жизнь. Если мы не вмешиваемся и даем самим пациентам решить, принимать ли лекарство, эти иные факторы могут повлиять на их решение, и разница в продолжительности жизни у тех, кто принимает и не принимает лекарство, больше не будет объясняться только этим. Например, представьте, что лекарство принимают только смертельно больные люди. Они определенно будут отличаться о тех, кто его не принимал, и сравнение двух групп будет отражать разницу в серьезности их болезни, а не эффект от лекарства. Однако, если заставлять пациентов принимать лекарство или отказываться от него, независимо от их изначального состояния, эта разница перестанет иметь значение и можно будет сделать обоснованное сравнение.

На языке математики мы записываем наблюдаемую частоту продолжительности жизни L у пациентов, которые добровольно приняли лекарство, как $P\,(L \mid D)$, и это стандартная условная вероятность, которая используется в учебниках по статистике. Это выражение подразумевает, что веро-

ятность P продолжительности жизни L допускается только в случае, если мы увидим, что пациент принимает лекарство D. Учтите, что $P(L \mid D)$ может резко отличаться от $P(L \mid do(D))$. Это разница между увиденным и сделанным фундаментальна, она объясняет, почему мы не считаем падение атмосферного давления причиной надвигающегося шторма. Если мы увидим, что падение атмосферного давление повышает вероятность шторма и заставим показания барометра измениться, мы, однако, никак не повлияем на эту вероятность.

Эта путаница между тем, что мы видим, и тем, что происходит, привела к изобилию парадоксов, и некоторые из них мы разберем в этой книге. Мир, лишенный $P(L \mid do(D))$ и управляемый исключительно $P(L \mid D)$, был бы действительно странным местом. Например, пациенты не ходили бы к врачу, чтобы избежать вероятности серьезно заболеть; города отказались бы от пожарных, чтобы сократить вероятность пожаров; врачи рекомендовали бы лекарства пациентам мужского и женского пола, но не пациентам, гендер которых неизвестен, и т. д. Трудно поверить, что менее трех десятилетий назад наука действовала в таком мире: оператора *do* не существовало.

Одним из главных достижений Революции Причинности стала возможность объяснить, как предсказать эффекты интервенции без ее осуществления. Это не было бы доступным, если бы, во-первых, мы не определили оператор *do*,

с помощью которого формулируется верный вопрос, и, во-вторых, не нашли бы способ моделировать его без реального вмешательства.

Когда интересующий нас научный вопрос подразумевает ретроспективное мышление, мы полагаемся на еще один тип причинного рассуждения – контрфактивное. Предположим, что Джо принял лекарство *D* и умер через месяц; нас интересует вопрос, могло ли лекарство вызвать его смерть. Чтобы разобраться в этом, нужно вообразить сценарий, при котором Джо уже собирался принять лекарство, но передумал. Выжил ли бы он?

И вновь скажем, что классическая статистика только обобщает данные, поэтому она не обеспечивает даже язык для ответа на такие вопросы. Наука о причинном анализе предоставляет систему обозначений, и, что важнее, предлагает решение. Как и в случае с эффектом интервенций (упомянутым выше), во многих ситуациях мы можем моделировать ретроспективное мышление человека с помощью алгоритма, который использует то, что мы знаем о наблюдаемом мире, и дает ответ о контрфактивном мире. Такая «алгоритмизация контрфактивного» – еще одна жемчужина Революции Причинности.

Контрфактивное рассуждение, основанное на «что, если», кажется ненаучным. Действительно, эмпирическое наблюдение не способно ни подтвердить, ни опровергнуть ответы на такие вопросы. Но наш ум постоянно делает весьма надеж-

ные и воспроизводимые суждения о том, что может быть или могло бы быть. Например, все мы понимаем, что, если бы петух не кричал этим утром, солнце все равно бы встало. Это согласие основано на том факте, что контрфактивные суждения – не игра воображения, а размышление о самой структуре нашей модели мира. Два человека, у которых одна и та же модель причинности, придут к одним и тем же контрфактивным суждениям.

Контрфактивные суждения – это строительные кирпичи этичного поведения и научной мысли. Способность размышлять о своих действиях в прошлом и предвидеть альтернативные сценария – это основа свободной воли и социальной ответственности. Алгоритмизация контрфактивных суждений открывает думающим машинам эту возможность, и теперь они могут разделить этот (доселе) исключительно человеческий способ осмыслять мир.

Я сознательно упомянул думающие машины в предыдущем абзаце. Я пришел к этой теме, когда занимался компьютерными науками, конкретно искусственным интеллектом, что обобщает две точки отправления для большинства из моих коллег, занятых причинным анализом. Во-первых, в мире искусственного интеллекта вы по-настоящему не понимаете тему до тех пор, пока не обучите ей робота. Вот почему вы увидите, что я неустанно, раз за разом подчеркиваю важность системы обозначений, языка, словаря и грамматики. Например, меня завораживает вопрос, в состоянии ли мы

выразить определенное утверждение на том или ином языке и следует ли это утверждение из других. Поразительно, сколько можно узнать, просто следуя грамматике научных высказываний! Мой акцент на язык также объясняется глубоким убеждением в том, что последний оформляет наши мысли. Нельзя ответить на вопрос, который вы не способны задать, и невозможно задать вопрос, для которого у вас нет слов. Изучая философию и компьютерные науки, я заинтересовался причинным анализом во многом потому, что мог с волнением наблюдать, как зреет и крепнет забытый когда-то язык науки.

Мой опыт в области машинного обучения тоже мотивировал меня изучать причинность. В конце 1980-х годов я осознал, что неспособность машин понять причинные отношения, вероятно, самое большое препятствие к тому, чтобы наделить их интеллектом человеческого уровня. В последней главе этой книге я вернусь к своим корням, и вместе мы исследуем, что значит Революция Причинности для искусственного интеллекта. Я полагаю, что сильный искусственный интеллект – достижимая цель, которой, к тому же не стоит бояться именно потому, что причинность – часть решения. Модуль причинного осмысления даст машинам способность размышлять над своими ошибками, выделять слабые места в своем программном обеспечении, функционировать как моральные сущности и естественно общаться с людьми о собственном выборе и намерениях.

Схема реальности

В нашу эпоху всем читателям, конечно, уже знакомы такие термины, как «знания», «информация», «интеллект» и «данные», хотя разница между ними или принцип их взаимодействия могут оставаться неясными. А теперь я предлагаю добавить в этот набор еще один термин – «причинная модель», после чего у читателей, вероятно, возникнет закономерный вопрос: не усложнит ли это ситуацию?

Не усложнит! Более того, этот термин свяжет ускользающие понятия «наука», «знания» и «данные» в конкретном и осмысленном контексте и позволит нам увидеть, как они работают вместе, чтобы дать ответы на сложные научные вопросы. На рис. 1. показана схема механизма причинного анализа, которая, возможно, адаптирует причинные умозаключения для будущего искусственного интеллекта. Важно понимать, что это не только проект для будущего, но и схема того, как причинные модели работают в науке уже сегодня и как они взаимодействуют с данными.

Механизм причинного анализа – это машина, в которую поступают три вида входных переменных – *допущения, запросы* и *данные* – и которая производит три типа выходных данных. Первая из входных переменных – решение «да/нет» о том, можно ли теоретически ответить на запрос в существующей причинной модели, если данные будут без-

ошибочными и неограниченными. Если ответ «да», то механизм причинного анализа произведет *оцениваемую величину*. Это математическая формула, которая считается рецептом для получения ответа из любых гипотетических данных, если они доступны. Наконец, после того как в механизм причинного анализа попадут данные, он использует этот рецепт, чтобы произвести действительную *оценку*. Подобная неопределенность отражает ограниченный объем данных, вероятные ошибки в измерениях или отсутствие информации.

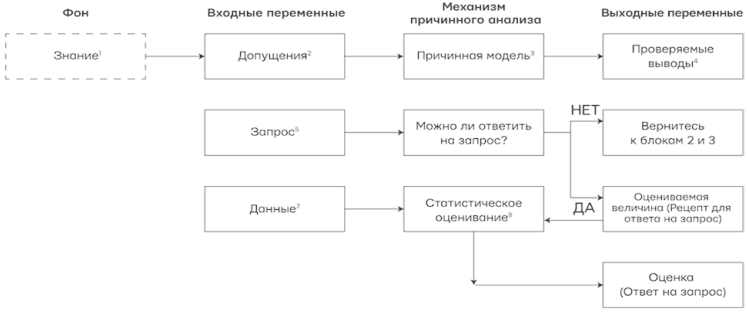

Рис. 1. Как механизм причинного анализа связывает данные со знанием причин, чтобы дать ответы на интересующие нас запросы. Блок, обозначенный пунктиром, не входит в механизм, но необходим для его построения. Также можно нарисовать стрелки от блоков 4 и 9 к блоку 1, но я решил сделать схему проще.

Чтобы объяснить схему подробнее, я пометил блоки цифрами от 1 до 9, и теперь прокомментирую их на примере запроса «Какой эффект лекарство D оказывает на продолжительность жизни *L*?»

1. «Знание» обозначает следы опыта, которые делающий умозаключения получил в прошлом. Это могут быть наблюдения из прошлого, действия в прошлом, а также образование и культурные традиции, признанные существенными для интересующего нас запроса. Пунктир вокруг «Знания» обозначает, что оно имеется в виду делающим умозаключения и не находит выражения в самой модели.

2. Научное исследование всегда требует упрощать допущения, т. е. утверждения, которые исследователь признает достойными, чтобы сформулировать их на основе доступного *знания*. Большая его часть остается подразумеваемой исследователем, и в модели запечатлены только допущения, которые получили формулировку и таким образом обнаружили себя. В принципе, их реально вычленить из самой модели, поэтому некоторые логики решили, что такая модель представляет собой всего лишь список допущений. Специалисты по компьютерным наукам делают здесь исключение, отмечая, что способ, избранный для представления допущений, в состоянии сильно повлиять на возможность правильно их сформулировать, сделать из них выводы и даже про-

должить или изменить их в свете новой убедительной информации.

3. Причинные модели записываются в разной форме. Это могут быть диаграммы причинности, структурные уравнения, логические утверждения и т. д. Я убежденный приверженец диаграмм причинности почти во всех случаях – прежде всего из-за их прозрачности, но также из-за конкретных ответов, которые они дают на многие вопросы, которые нам хотелось бы задать. Для этой диаграммы определение причинности будет простым, хотя и несколько метафорическим: переменная X – причина Y, если Y «слушает» X и приобретает значение, реагируя на то, что слышит. Например, если мы подозреваем, что продолжительность жизни пациента L «прислушивается» к тому, какое лекарство D было принято, то мы называем D причиной L и рисуем стрелку от D к L в диаграмме причинности. Естественно, ответ на наш вопрос о D и L, вероятно, зависит и от других переменных, которые тоже должны быть представлены на диаграмме вместе с их причинами и следствиями (здесь мы обозначим их совокупно как Z).

4. Эта практика слушания, предписанная путями в причинной модели, обычно приводит к наблюдаемым закономерностям или зависимостям в данных. Подобные закономерности называются проверяемыми выводами, потому что они могут быть использованы для проверки модели. Это утверждение вроде «Нет путей, соединяющих D и L», кото-

рое переводится в статистическое утверждение «D и L независимы», т. е. обнаружение D не влияет на вероятность L. Если данные противоречат этому выводу, то модель нужно пересмотреть. Чтобы это сделать, требуется еще один механизм, которые получает входные переменные из блоков 4 и 7 и вычисляет «степень пригодности», или степень, до которой *данные* совместимы с допущениями модели. Чтобы упростить диаграмму, я не стал показывать второй механизм на рис. 1.

5. Запросы, поступающие в механизм причинного анализа, – это научные вопросы, на которые мы хотим ответить. Их необходимо сформулировать, используя термины причинности. Скажем, что такое $P(L \mid do(D))$? Одно из главных достижений Революции Причинности состоит в том, что она сделала этот язык научно прозрачным и математически точным.

6. Оцениваемая величина – это статистическая величина, которая оценивается на основе данных. После оценки данных она в состоянии обоснованно представить ответ на наш запрос. Если записать ее как формулу вероятности, например $P(L \mid D, Z) \times P(Z)$, то фактически получишь рецепт, как ответить на причинный запрос с помощью имеющихся у нас данных, когда механизм причинного анализа подтвердит эту возможность.

Очень важно осознавать, что, в отличие от традиционной оценки в статистике, нынешняя модель причинности порой

не позволяет ответить на некоторые запросы, даже если какие-то данные уже собраны. Предположим, если наша модель покажет, что и D, и L зависят от третьей переменной Z (скажем, стадии болезни), и если у нас не будет способа измерить Z, то на запрос $P(L \mid do(D))$ нельзя будет получить ответ. В этом случае сбор данных окажется пустой тратой времени. Вместо этого придется вернуться назад и уточнить модель, либо добавив новые научные знания, которые позволят оценить Z, либо сделав допущения, которые все упростят (рискуя оказаться неправыми), например о том, что эффектом Z на D можно пренебречь.

7. Данные – это ингредиенты, которые используются в рецепте оцениваемой величины. Крайне важно осознавать, что данные абсолютно ничего не сообщают нам об отношениях причинности. Они обеспечивают нам значения, такие как $P(L \mid D)$ или $P(L \mid D, Z)$. Задача оцениваемой величины – показать, как «испечь» из этих статистических значений одну формулировку, которая с учетом модели будет логически эквивалентна запросу о причинности, скажем $P(L \mid do(D))$.

Обратите внимание, что само понятие оцениваемой величины и, более того, вся верхняя часть рис. 1 не существует в традиционных методах статистического анализа. Там оцениваемая величина и запрос совпадают. Так, если нам интересна доля тех, кто принимал лекарство D, среди людей с продолжительностью жизни L, мы просто записываем этот запрос как $P(D \mid L)$. То же значение и будет нашей оцени-

ваемой величиной. Оно уже определяет, какое соотношение данных надо оценить, и не требует никаких знаний о причинности. Именно поэтому некоторым статистикам по сей день чрезвычайно трудно понять, почему некоторые знания лежат за пределами статистики и почему одни только данные не могут заменить недостаток научного знания.

8. Оценка – то, что «выходит из печи». Однако она будет лишь приблизительной из-за еще одного свойства данных в реальном мире: они всегда относятся к ограниченной выборке из теоретически бесконечной популяции. В нашем текущем примере выборка состоит из пациентов, которых мы решили изучить. Даже если мы возьмем их произвольно, всегда останется некий шанс на то, что пропорции, которые мы определили, сделав измерения в выборке, не будут отражать пропорции в населении в целом. К счастью, статистика, как научная дисциплина, вооруженная продвинутыми приемами машинного обучения, дает нам великое множество способов справиться с этой неопределенностью: методы оценки максимальной вероятности, коэффициенты предрасположенности, интервалы доверия, критерии значимости и т. д. и т. п.

9. В итоге, если наша модель верна и если у нас достаточно данных, мы получаем ответ на запрос о причине, скажем такой: «Лекарство D повышает продолжительность жизни L у пациентов-диабетиков Z на 30 ± 20 %». Ура! Этот ответ добавит нам научных знаний (блок 1) и, если все пошло не так, как мы ожидали, обеспечит некоторые улучшения для

нашей модели причинности (блок 3).

На первый взгляд, эта диаграмма может показаться сложной, и вы, вероятно, задумаетесь, необходима ли она. Действительно, в повседневной жизни мы каким-то образом способны выносить суждения о причине, не проходя через такой сложный процесс и точно не обращаясь к математике вероятностей и пропорций. Одной нашей интуиции о причинности обычно достаточно, чтобы справиться с неопределенностью, с которой мы сталкиваемся каждый день дома или даже на работе. Но, если мы захотим научить тупого робота думать о причинах или раздвинуть границы научного знания, заходя в области, где уже не действует интуиция, тщательно структурированная процедура такого рода будет обязательной.

Я хочу особенно подчеркнуть роль данных в вышеописанном процессе. Для начала примите во внимание, что мы собираем данные, предварительно построив модель причинности, сформулировав научный запрос, на который хотим получить ответ и определив оцениваемую величину. Это противоречит вышеупомянутому традиционному для науки подходу, в котором даже не существует причинной модели.

Однако современная наука ставит новые вызовы перед теми, кто практикует рациональные умозаключения о причинах и следствиях. Хотя потребность в причинной модели в разных дисциплинах становится очевиднее с каждым днем,

многие исследователи, работающие над искусственным интеллектом, хотели бы избежать трудностей, связанных с созданием или приобретением причинной модели, и полагаться исключительно на данные во всех когнитивных задачах. Остается одна, в настоящий момент безмолвная надежда, что сами данные приведут нас к верным ответам, когда возникнут вопросы о причинности.

Я отношусь к этой тенденции с откровенным скепсисом, потому что знаю, насколько нечувствительны данные к причинам и следствиям. Например, информацию об эффекте действия или интервенции просто нельзя получить из необработанных данных, если они не собраны путем контролируемой экспериментальной манипуляции. В то же время, если у нас есть причинная модель, мы часто можем предсказать результат интервенции с помощью данных, к которым никто не прикасался.

Аргументы в пользу причинных моделей становятся еще более убедительными, когда мы пытаемся ответить на контрфактивные запросы, предположим: «Что бы произошло, если бы мы действовали по-другому?». Мы подробно обсудим контрфактивные запросы, потому что они представляют наибольшую сложность для любого искусственного интеллекта. Кроме того, развитие когнитивных навыков, сделавшее нас людьми, и сила воображения, сделавшие возможной науку, основаны именно на них. Также мы объясним, почему любой запрос о механизме, с помощью которого причины

вызывают следствия, – самый прототипический вопрос «Почему?» – на самом деле контрфактивный вопрос под прикрытием. Таким образом, если мы хотим, чтобы роботы начали отвечать на вопросы «Почему?» или хотя бы поняли, что они значат, их необходимо вооружить моделью причинности и научить отвечать на контрфактивные запросы, как показано на рис. 1.

Еще одно преимущество, которое есть у причинных моделей и отсутствует в интеллектуальном анализе данных и глубинном обучении, – это способность к адаптации. Отметим, что на рис. 1 оцениваемая величина определяется на базе одной только причинной модели – еще до изучения специфики данных. Благодаря этому механизм причинного анализа становится невероятно адаптивным, ведь оцениваемая величина в нем подойдет для любых данных и будет совместима с количественной моделью, какими бы ни были числовые зависимости между переменными.

Чтобы понять, почему эта способность к адаптации играет важную роль, сравните этот механизм с системой, которая пытается учиться, используя только данные. В этом примере речь пойдет о человеке, но в других случаях ей может быть алгоритм глубинного обучения или человек, использующий такой алгоритм. Так, наблюдая результат L у многих пациентов, которым давали лекарство D, исследовательница в состоянии предсказать, что пациент со свойством Z проживет L лет. Но теперь ее перевели в новую больницу в дру-

гой части города, где свойства популяции (диета, гигиена, стиль работы) оказались другими. Даже если эти новые свойства влияют только на числовые зависимости между зафиксированными переменными, ей все равно придется переучиваться и осваивать новую функцию предсказания. Это все, на что способна программа глубинного обучения – приспосабливать функцию к данным. Однако, если бы у исследовательницы была модель для действия лекарства и если бы ее причинная структура оставалась нетронутой в новом контексте, то оцениваемая величина, которую она получила во время обучения, не утратила бы актуальности. Ее можно было бы применить к новым данным и создать новую функцию предсказания.

Многие научные вопросы выглядят по-другому «сквозь линзу причинности», и мне очень понравилось возиться с этой линзой. В последние 25 лет ее эффект постоянно усиливается благодаря новым находкам и инструментам. Я надеюсь и верю, что читатели этой книги разделят мой восторг. Поэтому я хотел бы завершить это введение, анонсировав некоторые интересные моменты книги.

В главе 1 три ступени – наблюдение, интервенция и контрфактивные суждения – собраны в Лестницу Причинности, центральную метафору этой книги. Кроме того, здесь вы научитесь основам рассуждений с помощью диаграмм причинности, нашего главного инструмента моделирования, и встанете на путь профессионального овладения этим инструмен-

том. Более того, вы окажетесь далеко впереди многих поколений исследователей, которые пытались интерпретировать данные через линзу, непрозрачную для этой модели, и не знали о важнейших особенностях, которые открывает Лестница Причинности.

В главе 2 читатели найдут странную историю о том, как научная дисциплина статистика развила в себе слепоту к причинности и как это привело к далеко идущим последствиям для всех наук, зависящих от данных. Кроме того, в ней излагается история одного из величайших героев этой книги, генетика Сьюалла Райта, который в 1920-е годы нарисовал первые диаграммы причинности и долгие годы оставался одним из немногих ученых, осмелившихся воспринимать ее серьезно.

В главе 3 рассказывается равно любопытная история о том, как я обратился к причинности, работая над искусственным интеллектом – особенно над байесовскими сетями. Это был первый инструмент, который позволил компьютерам понимать «оттенки серого», и какое-то время я полагал, что они содержат главный ключ к искусственному интеллекту. К концу 1980-х годов я пришел к убеждению, что ошибался, и эта глава описывает мой путь от пророка до отступника. Тем не менее байесовские сети остаются очень важным инструментом для искусственного интеллекта и попрежнему во многом определяют математическое основания для диаграмм причинности. Помимо постепенного знаком-

ства с правилом Байеса и байесовскими методами рассуждения в контексте причинности, глава 3 представит увлекательные примеры того, как байесовские сети можно применить в реальной жизни.

Глава 4 рассказывает о главном вкладе статистики в причинный анализ – рандомизированном контролируемом исследовании (РКИ). С точки зрения причинности РКИ – это созданный человеком инструмент, позволяющий вскрыть запрос $P(L \mid do(D))$, возникший в природе. Главная его цель – отделить интересующие нас переменные (скажем, D и L) от других переменных (Z), которые в противном случае повлияли бы на обе предыдущие. Избавление от осложнений, вызванных такими неочевидными переменными, было проблемой в течение 100 лет. Эта глава показывает читателям удивительно простое ее решение, которое вы поймете за 10 минут, играючи проходя по путям в диаграмме.

Глава 5 повествует о поворотном моменте в истории причинности (и даже в истории всей науки), когда статистики столкнулись со сложностями, пытаясь выяснить, приводит ли курение к раку легких. Поскольку они не могли использовать свой любимый инструмент, РКИ, им было трудно прийти не только к единому выводу, но и к общему пониманию вопроса. Миллионы жизней оборвались или сократились из-за того, что ученым недоставало подходящего языка и методологии для ответов на вопросы о причинности.

Глава 6, надеюсь, даст читателям приятный повод от-

влечься от серьезных вопросов из главы 5. Это глава о парадоксах – Монти Холла, Симпсона, Берксона и др. Классические парадоксы такого рода можно рассматривать как занимательные головоломки, однако у них есть и серьезная сторона, которая видна особенно хорошо, если взглянуть на них с точки зрения причинности. Более того, почти все они отражают столкновения с причинной интуицией и таким образом обнажают анатомию этой интуиции. Словно канарейки в шахте, они сигнализировали ученым, что человеческая интуиция укоренена в причинной, а не статистической логике. Я полагаю, читателям понравится новый взгляд на любимые парадоксы.

Главы 7–9 наконец-то позволят читателю совершить увлекательный подъем по Лестнице Причинности. Мы начнем в главе 7 с интервенции, рассказывая, как я со студентами 20 лет пытался автоматизировать запросы типа *do*. В итоге нам удалось добиться успеха, и в этой главе объясняется, как устроен механизм причинного анализа», который дает ответ «да/нет», и что такое оцениваемая величина на рис. 1. Изучив этот механизм, читатель получит инструменты, которые позволят увидеть в диаграмме причинности некие структуры, обеспечивающие немедленный ответ на причинный запрос. Это «поправки черного входа», «поправки парадного входа» и инструментальные переменные – «рабочие лошадки» причинного анализа.

Глава 8 поднимет вас на вершину лестницы, поскольку в

ней рассматриваются контрфактивные суждения. Они считаются одной из необходимых составляющих причинности по меньшей мере с 1748 года, когда шотландский философ Дэвид Юм предложил для нее несколько искаженную дефиницию: «Мы можем определить причину как объект, за которым следует другой объект, если за всеми объектами, схожими с первым, следуют объекты, схожие со вторым. Или, другими словами, если бы не было первого объекта, второй бы не существовал». Дэвид Льюис, философ из Принстонского университета, умерший в 2001 году, указал, что на деле Юм дал не одно, а два определения: во-первых, регулярности (т. е. за причиной регулярно идет следствие) и, во-вторых, контрфактивности («если бы не было первого объекта…»). Хотя философы и ученые в основном обращали внимание на определение регулярности, Льюис предположил, что определение контрфактивности лучше сопрягается с человеческой интуицией: «Мы считаем причиной нечто, вызывающее перемену, и это перемена относительно того, что случилось бы без нее».

Читателей ждет приятный сюрприз: теперь мы можем отойти от научных дебатов и вычислить настоящее значение (или вероятность) для любого контрфактивного запроса – и неважно, насколько он изощрен. Особый интерес вызывают вопросы, связанные с необходимыми и достаточными причинами наблюдаемых событий. Например, насколько вероятно, что действие ответчика было неизбежной причиной

травмы истца? Насколько вероятно, что изменения климата, вызванные человеком, являются достаточной причиной аномальной жары?

Наконец, в главе 9 обсуждается тема медиации. Возможно, когда мы говорили о рисовании стрелок в диаграмме причинности, вы уже задавались вопросом, стоит ли провести стрелку от лекарства D к продолжительности жизни L, если лекарство влияет на продолжительность жизни только благодаря воздействию на артериальное давление Z (т. е. на посредника). Другими словами, будет ли эффект D, оказываемый на L, прямым или непрямым? И если наблюдаются оба эффекта, как оценить их относительную важность? Подобные вопросы не только представляют большой научный интерес, но и могут иметь практические последствия: если мы поймем механизм действия лекарства, то, скорее всего, сумеем разработать другие препараты с тем же эффектом, которые окажутся дешевле или будут иметь меньше побочных эффектов. Читателя порадует тот факт, что вечный поиск механизма медиации теперь сведен до упражнения в алгебре, и сегодня ученые используют новые инструменты из набора для работы с причинностью в решении подобных задач.

Глава 10 подводит книгу к завершению, возвращаясь к проблеме, которая изначально привела меня к причинности: как автоматизировать интеллект человеческого уровня (его порой называют сильным искусственным интеллектом). Я полагаю, что способность рассуждать о причинах абсолютно

необходима машинам, чтобы общаться с нами на нашем языке о политических мерах, экспериментах, объяснениях, теориях, сожалениях, ответственности, свободной воле и обязанностях – и в конечном счете принимать собственные этические решения.

Если бы я мог суммировать смысл этой книги в одной лаконичной и многозначительной фразе, она была бы такой: «Вы умнее ваших данных». Данные не понимают причин и следствий, а люди их понимают. Я надеюсь, что новая наука о причинном анализе позволит нам глубже осознать, как мы это делаем, ведь нет более эффективного способа понять себя, чем смоделировать себя. В эпоху компьютеров это новое знание также добавляет перспективу усилить наши врожденные способности, чтобы лучше постигать данные – как в больших, так и в малых объемах.

Глава 1. Лестница причинности

В начале…

Мне было, наверное, шесть или семь лет, когда я впервые прочел историю об Адаме и Еве в Эдемском саду. Мы с одноклассниками абсолютно не удивились капризным требованиям Бога, который запретил им есть плоды с древа познания. У божеств на все есть свои причины, думали мы. Но нас заинтриговал тот факт, что, когда Адам и Ева вкусили запретный плод, они, как и мы, стали осознавать свою наготу.

Когда мы стали подростками, наш интерес медленно сместился в сторону философских аспектов этой истории (израильские школьники читают Бытие несколько раз в год). Прежде всего нас взволновало, что возникновение человеческого знания было процессом не радостным, а болезненным – его сопровождали непослушание, вина и наказания. Некоторые спрашивали: имело ли смысл ради него отказываться от беззаботной жизни в Эдеме? И можно ли утверждать, что сельскохозяйственные и научные революции, которые случились после, стоили всех трудностей, войн и социальной несправедливости, неотъемлемых от современной жизни?

Не поймите меня неправильно: мы вовсе не были креационистами, и даже наши учителя были дарвинистами в ду-

ше. Однако мы знали, что автор, разыгравший эту историю по ролям, пытался ответить на самые насущные философские вопросы своего времени. Подобным образом мы ожидали, что она несет культурные отпечатки действительного процесса, в ходе которого *Homo sapiens* стал доминировать на нашей планете. Какой же в таком случае была последовательность шагов в этом скоростном процессе суперэволюции?

Интерес к таким вопросам угас, когда я на заре карьеры начал преподавать технические науки, но вдруг возродился в 1990-е годы, когда, работая над книгой «Причинность» (*Causality*), я познакомился с Лестницей Причинности.

Перечитывая Бытие в сотый раз, я заметил деталь, которая каким-то образом ускользала от моего внимания все эти годы. Когда Бог находит Адама, прячущегося в саду, он спрашивает: «… не ел ли ты от дерева, с которого Я запретил тебе есть?» И Адам отвечает: «… жена, которую Ты мне дал, она дала мне от дерева, и я ел». Бог спрашивает Еву: «… что ты это сделала?» Она отвечает: «… змей обольстил меня, и я ела».

Как мы знаем, Всемогущего не слишком впечатлили эти взаимные обвинения и он изгнал обоих из райского сада. И вот что я всегда пропускал до тех пор: Господь спросил: «Что?», а они ответили на вопрос «Почему?». Господь спрашивал о фактах, а они дали объяснения. Более того, оба бы-

ли полностью убеждены, что, если назвать причины, их действия будут каким-то образом выставлены в ином свете. Откуда они взяли эту мысль?

Для меня из этих деталей вытекают три глубоких вывода. Во-первых, еще на заре нашей эволюции мы, люди, осознали, что мир состоит не только из фактов (которые сегодня мы назвали бы данными); скорее, эти факты склеены вместе сложной сетью причинно-следственных отношений. Во-вторых, именно объяснения причин, а не сухие факты, составляют основу наших знаний и должны быть краеугольным камнем машинного интеллекта. Наконец, наш переход от обработчиков данных к создателям объяснений был не постепенным; потребовался скачок, который нуждался во внешнем толчке в виде необычного фрукта. Это в точности соответствовало тому, что я в теории наблюдал на Лестнице Причинности: ни одна машина не сможет извлечь объяснения из необработанных данных. Ей необходим толчок.

Если искать подтверждения для этих обобщений в науке об эволюции, то мы, конечно же, не найдем древа познания, но все же увидим важный необъяснимый переход. Сейчас мы понимаем, что люди произошли от обезьяноподобных предков за период от 5 до 6 миллионов лет и что такие постепенные эволюционные процессы вполне свойственны земной жизни. Но около 50 тысяч лет назад случилось нечто уникальное. Одни называют это Когнитивной Революцией, а другие (с некоторой иронией) – Великим Скачком. Лю-

ди приобрели способность менять окружающую среду и собственные возможности с принципиально иной скоростью.

Например, за миллионы лет эволюции у орлов и сов развилось потрясающее зрение, однако они так и не изобрели очки, микроскопы, телескопы или приборы ночного видения. Люди произвели эти чудеса в течение столетий. Я называю такой феномен суперэволюционным ускорением. Некоторые читатели могут возразить, утверждая, что я сравниваю абсолютно разные вещи – эволюцию и развитие техники, но в том-то и дело. Эволюция снабдила нас способностью внедрять технику в жизнь – дар, которым она не наделила орлов и сов, и здесь снова встает вопрос: почему? Как вычислительные навыки вдруг появились у людей, но не у орлов?

На этот счет было предложено много гипотез, но одна из них особенно тесно связана с идеей причинности. В книге «Sapiens: Краткая история человечества» Юваль Ной Харари постулирует, что способность наших предков воображать несуществующее стала ключевой, поскольку улучшила коммуникацию. До этого сдвига они могли доверять только людям из своей непосредственной семьи или племени. Потом их доверие распространилось на более крупные сообщества, объединенные общими фантазиями (например, верой в невидимых, но доступных воображению божеств, в загробную жизнь и в божественную сущность лидера) и ожиданиями. Согласитесь вы с гипотезой Харари или нет, но связь между воображением и причинными отношениями практи-

чески самоочевидна. Бесполезно говорить о причинах вещей, если вы не можете представить их последствий. Верно и обратное: нельзя утверждать, что Ева вынудила вас съесть плод с дерева, если вы не способны вообразить мир, в котором, вопреки фактам, она не дала вам яблока.

Но вернемся к нашим предкам *Homo sapiens*: новообретенная способность мыслить в категориях причинности позволила им делать много вещей эффективнее с помощью непростого процесса, который мы называем планированием. Представьте себе племя, которое готовится к охоте на мамонта. Что им потребуется для успеха? Признаться, я не лучший охотник на мамонтов, но, изучая думающие машины, я узнал одну вещь: думающая сущность (компьютер, пещерный человек или преподаватель вуза) способна выполнить задачу такого размаха, только если запланирует все заранее – решит, сколько охотников надо привлечь, оценит с учетом направления ветра, с какой стороны лучше приближаться к мамонту – в общем, вообразит и сравнит последствия нескольких стратегий охоты. Чтобы это сделать, думающая сущность должна обладать ментальной моделью реальности, сверяться с ней и манипулировать ей.

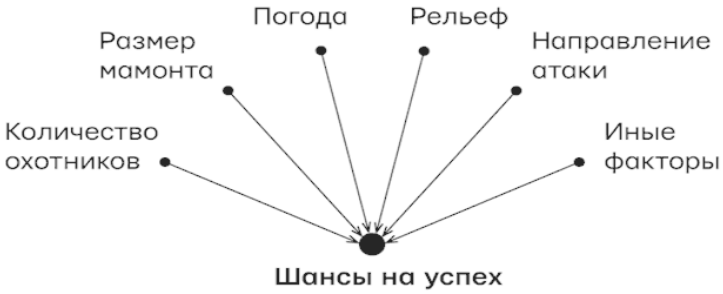

Рис. 2. Предполагаемые причины успеха в охоте на мамонта

Рисунок 2 показывает, как нарисовать такую модель в уме. Каждая точка на рисунке представляет собой причину успеха. Заметьте, что причин много и что ни одна из них не будет определяющей; т. е. мы не можем быть уверены, что большее число охотников обеспечит успех или что дождь гарантирует неудачу, однако эти факторы действительно влияют на вероятность успеха.

Ментальная модель – это арена, на которой работает воображение. Она позволяет экспериментировать с разными сценариями, внося изменения в конкретные места. Где-то в ментальной модели наших охотников был вспомогательный элемент, который позволял оценить эффект от числа участников. Когда они размышляли, стоит ли взять больше людей,

им не приходилось оценивать все остальные факторы с нуля. Они могли внести локальное изменение в модель, поставив «Охотники = 9» вместо «Охотники = 8», и снова оценить вероятность успеха. Этот модульный состав – основное свойство причинных моделей.

Я, конечно же, не хочу сказать, что первые люди рисовали себе модель, похожую на эту. Но когда мы пытаемся имитировать человеческую мысль на компьютере или даже когда хотим решить новые научные задачи, рисование картинок с конкретными точками и стрелками всегда исключительно полезно. Эти диаграммы причинности – вычислительная суть механизма причинного вывода, который я описал во вступлении.

Три уровня причинности

Возможно, к этому моменту я создал впечатление, что способность организовывать знания, деля их на причины и следствия, едина и мы приобрели ее сразу. На самом деле, исследуя машинное обучение, я узнал, что для изучения причинно-следственных связей необходимо овладеть когнитивными навыками по крайней мере на трех конкретных уровнях – видения, делания и воображения.

Первый навык, видение или наблюдение, подразумевает умение определять закономерности в окружающей среде. Он присутствует у многих животных и был у первых людей до Когнитивной Революции. Второй навык, делание, связан с умением предсказывать, какой эффект вызовут намеренные изменения в окружающей среде, и выбирать, какие изменения надо внести, чтобы получить желаемый результат. Очень немногие виды продемонстрировали элементы этого навыка. Использование инструментов, если это сознательные действия, а не случайность и не копирование предков, может свидетельствовать о переходе на этот следующий уровень. Но даже у пользователей инструментов не всегда есть «теория», которая говорит, почему инструмент работает и что делать, если он не работает. Для этого необходимо достичь уровня понимания, который допускает воображение. Именно этот третий уровень в первую очередь подготовил нас

к дальнейшим революциям в науке и сельском хозяйстве и резко преобразил воздействие нашего вида на планету.

Это я обосновать не могу, зато могу доказать математически, что три уровня фундаментально различны, и на каждом из них раскрываются способности, которых нет на предыдущих. Схема, которую я использую для демонстрации, восходит к Алану Тьюрингу, пионеру в исследовании искусственного интеллекта, предложившему классифицировать когнитивную систему, ориентируясь на вопросы, на которые она способна ответить. Такой подход оказался исключительно плодотворным, если говорить о причинности, потому что он позволяет избежать долгих и непродуктивных дискуссий о том, что именно представляет собой причинность, и сосредоточен на конкретном вопросе, на который реально ответить: что делает мыслитель, изучающий причинность? Или, если точнее, что может вычислить организм, имеющий модель причинности, тогда как организм, не имеющий модели причинности, это вычислить не в состоянии?

В то время как Тьюринг хотел создать бинарную классификацию, чтобы отличать человека от нечеловека, у нашей есть три уровня, соответствующих все более и более сложным причинным запросам. Используя эти критерии, можно собрать из запросов трех уровней одну Лестницу Причинности (рис. 3.) Мы будем еще не раз возвращаться к этой метафоре.

Давайте подробно рассмотрим каждую ее перекладину.

На первом уровне – ассоциаций – мы ищем повторяющиеся детали в наблюдениях. Этим занимается сова, которая наблюдает, как двигается крыса, и анализирует, где грызун окажется через секунду. Этим же занимается компьютерная программа для игры в го – она изучает базу данных с миллионами игр и может вычислить, какие ходы связаны с более высоким процентом выигрыша. Мы говорим, что одно событие связано с другим, если наблюдение одного изменения повышает вероятность увидеть другое.

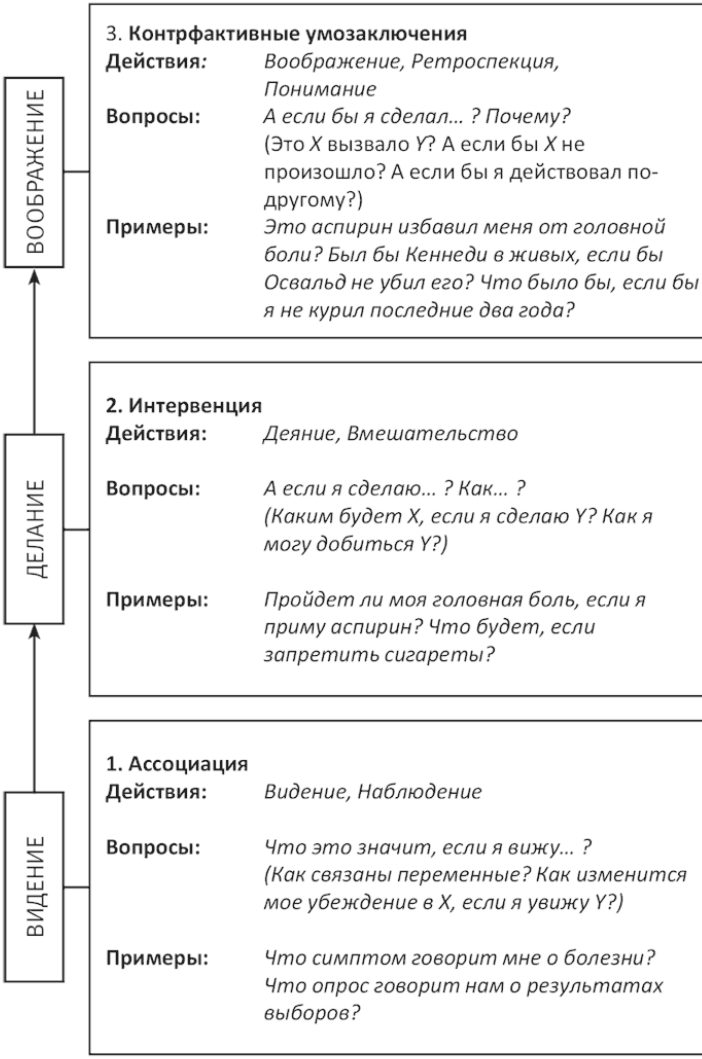
ВООБРАЖЕНИЕ
3. Контрфактивные умозаключения
Действия: Воображение, Ретроспекция, Понимание
Вопросы: А если бы я сделал... ? Почему? (Это X вызвало Y? А если бы X не произошло? А если бы я действовал по-другому?)
Примеры: Это аспирин избавил меня от головной боли? Был бы Кеннеди в живых, если бы Освальд не убил его? Что было бы, если бы я не курил последние два года?
ДЕЛАНИЕ
2. Интервенция
Действия: Деяние, Вмешательство
Вопросы: А если я сделаю... ? Как... ? (Каким будет X, если я сделаю Y? Как я могу добиться Y?)
Примеры: Пройдет ли моя головная боль, если я приму аспирин? Что будет, если запретить сигареты?
ВИДЕНИЕ
1. Ассоциация
Действия: Видение, Наблюдение
Вопросы: Что это значит, если я вижу... ? (Как связаны переменные? Как изменится мое убеждение в X, если я увижу Y?)
Примеры: Что симптом говорит мне о болезни? Что опрос говорит нам о результатах выборов?

Рис. 3. Лестница Причинности с представляющими ее организмами на каждом уровне. Большинство животных, так же как и сегодняшние обучающиеся машины, находятся на первой перекладине – они учатся по ассоциации. Пользователи инструментов вроде первых людей находятся на второй перекладине – если действуют по плану, а не просто имитируют. Кроме того, на этом уровне можно ставить эксперименты, чтобы узнать, какой эффект дает интервенция. Предположительно именно так младенцы получают большинство знаний о причинности. Те же, кто учится с помощью контрфактивных рассуждений, находятся на верхней перекладине и могут вообразить несуществующие миры и назвать причины для наблюдаемых феноменов.

Первая перекладина лестницы подразумевает предсказания, основанные на пассивных наблюдениях. Ее характеризует вопрос: «Что, если я увижу…?» Например, представьте директора по маркетингу в универмаге, который спрашивает: «Какова вероятность, что потребитель, который купил зубную пасту, также приобретет зубную нить?» Такие вопросы – самая суть статистики, и на них отвечают прежде всего, собирая и анализируя данные. В нашем случае на этот вопрос получится ответить, взяв данные о покупательском поведении всех клиентов, выбрав тех, кто купил зубную пасту, и, сосредоточившись на последней группе, вычислить долю тех, кто приобрел еще и зубную нить. Эта пропорция, также

известная как условная вероятность, измеряет (для больших объемов данных) степень связи между покупкой пасты и покупкой зубной нити. Мы можем записать это в символах как *P (зубная нить | зубная паста)*. *P* обозначает вероятность, вертикальная линия – «при условии, что вы видите».

Статистики предложили много изощренных методов, которые позволяют сократить большой объем данных и выявить связи между переменными. Корреляция или регрессия – типичная мера взаимосвязи, которая часто упоминается в этой книге. Чтобы увидеть ее, необходимо провести линию, ориентируясь на распределение единиц наблюдения, и продолжить ее уклон. Некоторые связи имеют очевидную интерпретацию с точки зрения причинности; другие могут ее не иметь. Но одна только статистика не скажет нам, что причина, а что следствие – зубная паста или зубная нить. С точки зрения менеджера по продажам это может не иметь особого значения. Точные предсказания не нуждаются в хороших объяснениях. Сова отлично охотится, не понимая, почему крыса всегда движется из точки *A* в точку *B*.

Некоторые читатели могут быть удивлены тем, что я разместил обучающиеся машины наших дней прямо на первой перекладине Лестницы Причинности – рядом с мудрой совой. Такое ощущение, что почти каждый день мы слышим о стремительном прогрессе систем машинного обучения – о самоуправляемых автомобилях, системах распознавания речи и, особенно в последнее время, об алгоритмах глубинно-

го обучения (или глубинных нейросетях). Как же они могут до сих пор оставаться на первом уровне?

Успехи глубинного обучения стали по-настоящему примечательными и оказались сюрпризом для многих из нас. В то же время глубинное обучение оказалось успешным в основном потому, что показало: определенные вопросы или задания, которые мы считали трудными, на самом деле не являются таковыми. Оно не коснулось по-настоящему сложных вопросов, которые до сих пор не дают нам создать искусственный интеллект, подобный человеческому. В результате общественность верит, что машины с «сильным ИИ», которые думают, как человек, вот-вот появятся или, возможно, уже появились. В реальности это максимально далеко от правды. Я полностью согласен с Гэри Маркусом, нейроученым из Нью-Йоркского университета, который недавно писал в «Нью-Йорк таймс» о том, что сфера искусственного интеллекта «полнится микрооткрытиями», которых хватает для хороших пресс-релизов, но машины все еще огорчительно далеки от познания, подобного человеческому. Мой коллега Эднан Дарвиш, специалист по компьютерным наукам из Калифорнийского университета в Лос-Анджелесе, назвал свою программную статью «Интеллект как у человека или способности как у животных?» и, я думаю, очень точно поставил в ней интересующий нас вопрос. Сильный искусственный интеллект нужен для того, чтобы производить машины с интеллектом, подобным человеческому, которые

будут способны общаться с людьми и направлять их. В то же время глубинное обучение дает нам машины с действительно впечатляющими способностями, но без интеллекта. Разница здесь глубокая, и ее причина – отсутствие модели реальности.

Точно так же, как 30 лет назад, программы машинного обучения (включая программы с глубинными нейросетями) практически всегда действуют в режиме ассоциаций. Они используют поток наблюдений, к которым пытаются приспособить функцию, по существу как статистик, который старается увидеть линию в скоплении точек – единиц информации. Глубинные нейросети повышают сложность подобранной функции, добавляя много слоев, но процесс подбора до сих пор базируется на необработанных данных. Чем больше данных используется, тем выше становится точность, но «суперэволюционного ускорения» не происходит. Если, например, программисты беспилотной машины захотят, чтобы она по-разному реагировала на новые ситуации, им придется быстро добавить эти новые реакции. Машина сама не поймет, что пешеход с бутылкой виски в руке, вероятно, по-своему отреагирует на сигнал. Это отсутствие гибкости и приспособляемости неизбежно для любой системы, которая работает на первом уровне нашей Лестницы Причинности.

Мы переходим на следующую ступень запросов о причинности, когда начинаем менять мир. Обычный вопрос для этого уровня будет таким: «Как изменятся продажи зубной

нити, если удвоить стоимость зубной пасты?». Это уже требует нового вида знаний, которого нет в наших данных, обнаруженных на втором уровне Лестницы Причинности – интервенции.

Интервенция стоит выше ассоциации, потому что подразумевает не только наблюдение, но и изменение. Когда мы видим дым и когда дымим сами, это подразумевает совершенно разное представление о вероятности пожара. На вопросы об интервенции нельзя ответить с помощью пассивно собранных данных, и неважно, насколько велик их объем или насколько глубока нейронная сеть. Для многих ученых стала настоящим ударом информация о том, что никакие методы, известные из статистики, не позволяют даже выразить простой вопрос, например «Что будет, если мы удвоим цену?», не говоря уже о его решении. Я знаю это, поскольку много раз помогал им подняться на следующую перекладину лестницы.

Почему нельзя ответить на вопрос о зубной нити просто при помощи наблюдения? Ведь можно заглянуть в нашу обширную базу данных о предыдущих покупках, посмотреть, что было раньше, когда зубная паста стоила в два раза больше? Причина в том, что в предыдущих случаях цена могла быть выше по другим причинам. Предположим, товара осталось немного и всем остальным магазинам тоже пришлось повысить цены. Но теперь вы размышляете о намеренном вмешательстве, после которого установится новая цена,

независимо от условий на рынке. Результат может сильно отличаться от предыдущего, когда покупатель не мог купить товар по более выгодной цене в других местах. Если бы у вас были данные об условиях на рынке в других ситуациях, вероятно, вы смогли бы предсказать все это лучше, но какие данные нужны? И как это выяснить? Наука о причинном выводе позволяет нам отвечать именно на эти вопросы.

Непосредственный способ предсказать результат интервенции – провести с ней эксперимент в тщательно контролируемых условиях. Компании, работающие с большими данными, такие как «Фейсбук», знают об этом и постоянно ставят эксперименты, чтобы посмотреть, что случится, если по-другому разместить элементы на экране или показать клиенту новую подсказку (либо даже новую цену).

Еще интереснее тот факт, что успешные предсказания об эффекте интервенции иногда можно сделать даже без эксперимента, хотя это не так широко известно, и даже в Кремниевой долине. Предположим, менеджер по продажам создает модель потребительского поведения и учитывает в ней ситуацию на рынке. Если данных обо всех факторах не имеется, вероятно, получится подставить достаточно суррогатных ключей и сделать прогноз. Сильная и точная причинная модель позволит использовать данные с первого уровня (наблюдения), чтобы ответить на запросы со второго уровня (об интервенции). Без причинной модели нельзя перейти с первой перекладины Лестницы на вторую. Вот почему системы

глубинного обучения (если в них используются только данные с первой перекладины и нет причинной модели) никогда не смогут отвечать на вопросы об интервенции, по определению нарушающие правила среды, в которой обучалась машина.

Как иллюстрируют все эти примеры, главный вопрос на второй перекладине Лестницы Причинности – «Что, если мы…?». Что произойдет, если мы *изменим* среду? Можно написать запрос P (*нить* | do (*зубная паста*)), чтобы узнать, какова вероятность продать зубную нить по определенной цене, если мы будем продавать зубную пасту по другой цене.

Еще один популярный вопрос на этом уровне причинности – «Как?» Это родственник вопроса «Что, если мы…?». Скажем, менеджер говорит нам, что на складе слишком много зубной пасты. Он спрашивает: «Как нам ее продать?», т. е. какую цену лучше на нее назначить. И снова вопрос относится к интервенции, которую нужно совершить в уме, прежде чем решить, стоит ли осуществлять ее в реальной жизни и как это осуществить. Здесь требуется модель причинности.

В повседневной жизни мы постоянно совершаем интервенции, хотя обычно не называем их таким замысловатым термином. Предположим, принимая аспирин, чтобы избавиться от головной боли, мы вмешиваемся в одну переменную (количество аспирина в нашем организме), чтобы повлиять на другую (состояние головной боли). Если наш причинный взгляд на аспирин верен, то переменная результата

отреагирует, изменившись с «головной боли» на «отсутствие головной боли».

Хотя рассуждения об интервенциях – важный уровень на Лестнице Причинности, все же они не отвечают на все интересующие нас вопросы. Можно задуматься: головная боль прошла, но почему? Помог аспирин? Или что-то из еды? Хорошие новости, которые я услышал? Эти вопросы приводят нас на верхний уровень Лестницы Причинности – уровень контрфактивных суждений, потому что для ответа на них нужно вернуться в прошлое, изменить историю и спросить себя: что случилось бы, если бы я не принял аспирин? Никакой эксперимент в мире не может отменить лечение человеку, который уже исцелился, и не позволит сравнить два исхода, поэтому необходимо применить совершенно новый вид знания.

Контрфактивные суждения находятся в особенно проблематичных отношениях с данными, потому что последние по определению относятся к фактам. Они не могут сообщить нам, что случится в контрфактивном или воображаемом мире, где некоторые наблюдаемые факты резко отвергаются. Но все же человеческий разум производит логические рассуждения такого рода – постоянно и с высокой надежностью. Это сделала Ева, когда обозначила причину своих действий: «Змей обольстил меня». Такая способность больше всего отличает человеческий интеллект от интеллекта животного, равно как и от невосприимчивых к подобным моделям вер-

сий ИИ и обучающихся машин.

Вероятно, вам не верится, что наука способна сделать полезные заключения в духе «а что, если» о мирах, которые не существуют, и о вещах, которые не происходили. Однако этим она и занимается – и занималась всегда. Законы физики можно рассматривать как контрфактивные утверждения, например: «Если бы вес этой спирали удвоился, ее длина тоже удвоилась бы» (закон Гука). Это утверждение, конечно, поддерживается изобилием экспериментальных подтверждений (второго уровня), полученных с помощью сотен спиралей в десятках лабораторий в тысячах случаев. Однако, поскольку утверждение нарекли законом, физики интерпретируют его как функциональную зависимость, которая управляет конкретной спиралью в конкретный момент при гипотетических значениях веса. Все эти разные миры, где вес составляет x кг, а длина спирали – L_X см, рассматриваются как объективно известные и одновременно действующие, хотя на самом деле существует только один из них.

Если вернуться к примеру с зубной пастой, то вопрос на верхнем уровне будет таким: какова вероятность, что покупатель зубной пасты все равно купил бы ее, если бы мы удвоили цену? Мы сравниваем реальный мир (в котором знаем, что покупатель приобрел зубную пасту по текущей цене) с воображаемым миром (где цена вдвое выше).

Если иметь причинную модель, которая способна ответить на контрфактивные вопросы, преимущества будут

огромными. Если понять причины грубой ошибки, в будущем можно будет принять меры, которые позволят все скорректировать. Если понять, почему лекарство помогло одним, но не помогло другим, получится открыть новые способы лечить болезнь. Отвечая на вопрос, как сложились бы события, если бы что-то пошло по-другому, мы извлечем уроки из истории и опыта других людей, и, кажется, ни один другой вид на это не способен. Неудивительно, что греческий философ Демокрит (около 460 – около 370 года до н. э.) сказал: «Я предпочел бы найти одну-единственную причину, чем стать персидским царем».

Расположение контрфактивных суждений на верхнем уровне Лестницы Причинности объясняет, почему я придаю им такое значение как ключевому моменту в эволюции человеческого создания. Я полностью согласен с Ювалем Харари в том, что описание воображаемых существ было демонстрацией новой способности, которую он называет Когнитивной Революцией. Ее классический пример – статуэтка человекольва, найденная в пещере Штадель в юго-западной Германии, которая сейчас хранится в Ульмском музее. Человеколев, созданный около 40 тысяч лет назад, представляет собой химеру, наполовину льва и наполовину человека, вырезанную из бивня мамонта.

Мы не знаем, кто создал человекольва и с какой целью это было сделано, но мы все же знаем, что это были анатомически современные люди и что это знаменует разрыв со все-

ми искусствами и ремеслами, практиковавшимися прежде. Раньше люди изготовляли инструменты и предметы фигуративного искусства – от бусин до флейт, наконечников копий и элегантных статуэток лошадей и прочих животных. Человеколев имеет иную природу – это творение чистого воображения.

Демонстрируя нашу новообретенную способность воображать вещи, которые никогда не существовали, человеколев является предшественником всех философских теорий, научных открытий и технических инноваций – от микроскопов до самолетов и компьютеров. Все они сначала появились в чьем-то воображении, а уже потом воплотились в физическом мире.

Этот скачок когнитивных возможностей был таким же глубоким и важным для нашего вида, как и все анатомические изменения, которые сделали нас людьми. В течение 10 тысяч лет после создания человекольва все иные виды рода *Homo* (кроме очень изолированного географически человека флоресского) вымерли. А люди продолжили менять естественный мир с невероятной скоростью, используя воображение, чтобы выжить, приспособиться и в итоге доминировать. Преимущество, которое мы получили, воображая контрфактивные ситуации, было тем же, что и сегодня: оно давало гибкость, способность размышлять и совершенствоваться на основе действий в прошлом и, что, вероятно, еще важнее, готовность брать на себя ответственность за дей-

ствия в прошлом и будущем.

Как показано на рис. 3, для третьего уровня Лестницы Причинности характерны запросы вроде «Что было бы, если бы я сделал…?» и «Почему?». Оба подразумевают сравнение наблюдаемого мира с контрфактивным миром. Эксперименты сами по себе не позволяют отвечать на такие вопросы. В то время как на первом уровне мы имеем дело с наблюдаемом миром, а на втором уровне – с дивным новым миром, который можно увидеть, на третьем уровне идет взаимодействие с миром, который увидеть нельзя (потому что он противоречит наблюдаемому). Чтобы преодолеть этот разрыв, необходима модель причинного процесса, который иногда называют теорией или (когда мы невероятно уверены в себе) законом природы. Короче говоря, нам необходимо понимание. Это, конечно же, святой Грааль любой науки – разработка теории, которая позволит нам предсказать, что случится в ситуациях, которые мы даже не предвидели. Но дело заходит еще дальше: присутствие таких законов позволяет нам выборочно нарушать их, чтобы создать мир, который противоречит нашему. В следующем разделе мы рассмотрим такие нарушения на практике.

Мини-тест Тьюринга

В 1950 году Алан Тьюринг задался вопросом, что это значит: компьютер, думающий как человек. Он предложил практический тест под названием «Игра в имитацию», но исследователи искусственного интеллекта с тех пор зовут его исключительно тестом Тьюринга. Во всех практических отношениях компьютер достоин считаться думающей машиной, если обычный человек, который общается с ним при помощи клавиатуры, не догадается, с кем он разговаривает – с другим человеком или с компьютером. Тьюринг был горячо уверен в том, что это абсолютно достижимо. Он писал: «Я верю, что примерно через 50 лет можно будет так хорошо программировать компьютеры для игры в имитацию, что после пяти минут вопросов и ответов у среднего собеседника будет не более 70 %-ного шанса сделать правильный выбор».

Предсказание Тьюринга оказалось немного неточным. Ежегодно самый похожий на человека чатбот в мире борется за премию Лёбнера: за программу, которая сумеет обмануть всех четырех судей, притворяясь человеком, полагается золотая медаль и 100 тысяч долларов. В 2015 году, спустя 25 лет с начала соревнований, ни одной программе не удалось обмануть не то что всех судей, но даже и половину.

Тьюринг не просто разработал игру в имитацию, он также предложил стратегию, чтобы пройти тест. «Что, если разра-

ботать программу, симулирующую не разум взрослого человека, а ум ребенка?» – спросил он. Если это сделать, можно было бы обучить ее так, как мы обучаем детей, – и вуаля! Через 20 лет (или меньше, учитывая более высокую скорость компьютера) мы получим искусственный интеллект. «Можно предположить, что ум ребенка подобен тетради, которую покупают в канцелярском магазине, – писал он. – Совсем небольшой механизм и много пустых страниц». Здесь он ошибался: мозг ребенка богат механизмами и заранее загруженными шаблонами.

И все же я думаю, что в чем-то Тьюринг прав. Скорее всего, у нас не получится произвести интеллект, подобный человеческому, пока мы не создадим интеллект, схожий с детским, и главным компонентом этого интеллекта будет владение причинно-следственными связями.

Как же машины могут получить знания о причинно-следственных связях? Это и по сей день остается важнейшим вызовом, который, несомненно, относится к замысловатым сочетаниям данных, поступающих из активных экспериментов, пассивного наблюдения и (не в последней степени) самого программиста, что во многом похоже на входящую информацию, которую получает ребенок, только эволюцию, родителей и товарищей заменяет программист.

Тем не менее ответим на несколько менее амбициозный вопрос: как машины (и люди) могли бы представить знания о причинно-следственных связях таким образом, чтобы быст-

ро получать доступ к нужной информации, правильно отвечать на вопросы и делать это с такой же легкостью, с какой это получается у трехлетнего ребенка? На самом деле таков главный вопрос, который мы рассмотрим в этой книге.

Я называю это мини-тестом Тьюринга. Идея здесь в том, чтобы взять простую историю, каким-то образом закодировать ее на машине, а потом проверить, сможет ли она правильно ответить на вопросы о причинно-следственных связях, на которые способен ответить человек. Это мини-тест по двум причинам. Во-первых, потому что он сведен к рассуждениям о причинах и следствиях, что исключает остальные аспекты человеческого интеллекта, такие как общая картина мира и естественный язык. Во-вторых, мы позволяем конкурсанту закодировать историю в виде любого удобного представления и освобождаем машину от задачи извлечь историю из собственного опыта. Проходить этот мини-тест стало задачей всей моей жизни – я делаю это сознательно последние 25 лет и делал бессознательно раньше.

Очевидно, готовясь к мини-тесту Тьюринга, мы должны сначала ответить на вопрос о репрезентации, а уже потом – об усвоении информации. Без репрезентации мы не знали бы, как хранить данные для использования в будущем. Даже если бы мы могли дать роботу манипулировать окружающей средой по его желанию, любая информация, полученная таким образом, забылась бы, если бы роботу не дали шаблон, чтобы закодировать результаты этих манипуляций. Важней-

шим вкладом ИИ в исследование познания стала парадигма «Сначала репрезентация – потом усвоение». Часто поиск хорошей репрезентации приводил к ценным находкам о том, как стоит получать знания – и из данных, и от программиста.

Когда я описываю мини-тест Тьюринга, в ответ мне обычно утверждают, что его легко пройти с помощью обмана. Например, можно взять список всех вероятных вопросов, сохранить правильные ответы, а потом привести их по памяти, когда вас спросят. И тогда не будет способа отличить машину, в которой всего лишь хранится список вопросов и ответов, от машины, которая отвечает так же, как мы с вами, т. е. понимает вопрос и производит ответ, используя ментальную модель причинности. И что же докажет мини-тест Тьюринга, если жульничать так просто?

Философ Джон Сёрл в 1980 году описал эту возможность обмана с помощью мысленного эксперимента под названием «Китайская комната». Он подверг сомнению утверждение Тьюринга о том, что способность сымитировать интеллект равна обладанию им. С аргументом Сёрла есть только одна проблема: обмануть тест нелегко, более того, это нереально. Даже при ограниченном наборе переменных количество вероятных вопросов растет астрономически. Скажем, у нас есть 10 каузальных переменных и каждая из них может иметь два значения (0 или 1). Мы способны задать около 30 миллионов предполагаемых запросов, например: «Какова вероятность, что результат будет равен 1, если мы *увидим*

Конец ознакомительного фрагмента.